University of Tennessee, Knoxville
Trace: Tennessee Research and Creative Exchange

Doctoral Dissertations | Graduate School

8-2006

Integrated Computational and Experimental Platform for Characterizing Protein Isoforms and PTMs in Microbial Systems by Top-Down FT-ICR Mass Spectrometry

Heather Marie Connelly
University of Tennessee - Knoxville

Recommended Citation

Connelly, Heather Marie, "Integrated Computational and Experimental Platform for Characterizing Protein Isoforms and PTMs in Microbial Systems by Top-Down FT-ICR Mass Spectrometry. " PhD diss., University of Tennessee, 2006.
https://trace.tennessee.edu/utk_graddiss/1656

This Dissertation is brought to you for free and open access by the Graduate School at Trace: Tennessee Research and Creative Exchange. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of Trace: Tennessee Research and Creative Exchange. For more information, please contact trace@utk.edu.

To the Graduate Council:

I am submitting herewith a dissertation written by Heather Marie Connelly entitled "Integrated Computational and Experimental Platform for Characterizing Protein Isoforms and PTMs in Microbial Systems by Top-Down FT-ICR Mass Spectrometry." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

Robert L. Hettich, Major Professor

We have read this dissertation and recommend its acceptance:

Dale A. Pelletier, Gregory B. Hurst, Cynthia B. Peterson, Frank W. Larimer

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:

I am submitting herewith a dissertation written by Heather Marie Connelly entitled "Integrated Computational and Experimental Platform for Characterizing Protein Isoforms and PTMs in Microbial Systems by Top-Down FT-ICR Mass Spectrometry". I have examined the final electronic copy of this dissertation for form and content and recommended that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

Robert L. Hettich
Major Professor

We have read this dissertation
and recommended its acceptance:

Dale A. Pelletier

Gregory B. Hurst

Cynthia B. Peterson

Frank W. Larimer

Accepted for the Council:

Anne Mayhew
Vice Chancellor and
Dean of Graduate Studies

"Original signatures are on file with official student records."

Integrated Computational and Experimental Platform for Characterizing Protein Isoforms and PTMs in Microbial Systems by Top-Down FT-ICR Mass Spectrometry

A Dissertation Presented for the Doctor of Philosophy Degree
The University of Tennessee, Knoxville

Heather Marie Connelly
August, 2006

DEDICATION

I dedicate this dissertation to all those who have inspired and stood behind me. To my family Randy, Gale, Vicki and Nathan Connelly for believing I could accomplish my goals even when I sometimes doubted. To Honeydew, for always being happy to see and be with me, even when I had a really bad day. Finally, to Chris, for always providing support and an escape when I needed one.

ACKNOWLEDGMENT

I would like to thank the many people who assisted me in the completion of the research presented in this dissertation. I would first like to thank my graduate advisor Dr. Robert Hettich for guidance in my doctoral studies. I would like to thank Dr. Leos Kral for first introducing me to the wonderful world of biological research and always being there for advice when I needed it. I would like to thank Dr. Gregory Hurst, Dr. Dale Pelletier, Dr. Frank Larimer and Dr. Cynthia Peterson for serving on my doctoral committee and always providing feedback and biological insight.

I would like to thank all staff of the Organic and Biological Mass Spectrometry Group, including Dr. Doug Goeringer, Dr. Gary Van Berkel, Dr. Gregory B. Hurst, and Dr. Hayes McDonald for always taking the time to teach and assist in any way they could. I sincerely thank Becky R. Maggard of the Organic and Biological Mass Spectrometry Group for secretarial assistance in the preparation of many of the manuscripts that make up this dissertation and for assistance with the dissertation as a whole. I would like to thank all of the students and post-docs of the Organic and Biological Mass Spectrometry Group especially Vilmos Kertezs for working with me on software that helped make this dissertation possible. A special thank you goes to my mom, Gale, for patiently proofreading and editing this entire dissertation.

I would like to thank Dr. Giddings, and her lab, for collaborative efforts on the *E. coli* antibiotic resistant ribosomal protein project. I would also like to thank Patricia Lankford, Dr. Dale Pelletier, and Tse-Yuan Lu for collaborative efforts on the *Rhodopseudomonas palustris* GlnK project, as well as the top-down proteom project.

Lastly, I would like to acknowledge support from the University of Tennessee (Knoxville)-ORNL Graduate School of Genome Science and Technology. Much of the research presented here was funded by the U.S. Department of Energy (Office of Biological and Environmental Research, Office of Science) grants from the Genomes To Life and Microbial Genome Programs. Without continued financial support from the three institutes, none of this research would have been possible.

ABSTRACT

The goals of this dissertation research were to develop an integrated computational and experimental platform for characterizing protein isoforms and post translational modifications (PTMs) in microbial systems by top-down FT-ICR mass spectrometry. To accomplish this goal, we employed methodologies of microbial growth, intact protein and protein complex extractions, followed by sample preparation and then progressed to identification of the instrumentation needed to integrate the top-down and bottom-up proteomics methodologies used in these studies. Emphasis is placed on the development of integrated top-down and bottom-up informatics and the challenges faced in the integration of these two large mass spectrometry data sets and extraction of relevant biological data. We then illustrate how top-down and bottom-up methods can be applied to the analysis of complex protein mixtures, protein complexes, and microbial proteomes. Through the work of this dissertation we have contributed to the advancement of top-down proteomics by providing an experimental platform which will aid in the analysis of intact proteins and their associated PTMs and isoforms, as well as providing a computational method that allows for the integration of top-down and bottom-up data sets.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

AAC	Amino acid composition
AMT	Accurate mass tag
BCA	Bicinchoninic acid solution
CAD	Collisional activated dissociation
Capp	ES flow rates over 1 ul/min with LC
CNBr	Cyanogen bromide
COG	Cluster of orthologous groups
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
EM	Electron multiplier
ES	Electrospray ionization
ESI	Electrospray ionization
FA	Formic acid
FAB	Fast atom bombardment
FISH	Fluorescent *in-situ* hybridization
FPLC	Fast protein liquid chromatography
FT-ICR	Fourier Transform Ion Cyclotron Resonance
FT-MS	Fourier transform mass spectrometry
HFIP	Hexafluroisopropanol
HPLC	High performance liquid chromatography
i.d.	Internal diameter
JGI	Joint Genome Institute
KEGG	Kyoto Encyclopedia of Genes and Genomes
LC-MS	Liquid chromatography-mass spectrometry
LC-MS/MS	Liquid chromatography-tandem mass spectrometry
LCQ	Thermo Finnigan ES quadrupole ion trap
LTQ	Thermo Finnigan ES linear ion trap
MALDI	Matrix assisted laser desorption ionization
MASPIC	DBDigger scorer
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
MSAD	Multipole storage assisted dissociation
MW	Molecular weight
Nano	ES flow rates less than 1 ul/min with LC
ORF	Open reading frame
ORNL	Oak Ridge National Laboratory
PCR	Polymerase chain reaction
PMF	Peptide mass fingerprint
PPM	Parts per million
PTM	Post translational modification
QIT	Quadrupole ion trap
RP	Reverse phase
RPLC	Reverse phase liquid chromatography

SAX	Strong anion exchange
SCX	Strong cation exchange
SDS-PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
SORI-CAD	sustained off resonance irradiated collisional activated dissociation
SmR	Streptomycin resistant ecoli
SmRC	Streptomycin resistant compensated *E.coli*
TACT	Charge state determination software
TAP	Tandem affinity purification
TCA	Trichloroacetic acid
TFA	Trifluoroacetic acid
TIC	Total ion chromatogram
TIGR	The Institute for Genome Research
TOF	Time of flight
WT	Wild-type
Xcorr	SEQUEST cross-correlation score
2D-PAGE	Two-dimensional polyacrylamide gel electrophoresis

Chapter 1

Introduction to the Analysis of Intact Proteins and PTMs in Microbial Systems by Mass Spectrometry

Some of the text presented below has been published as Nathan C. VerBerkmoes, Heather M. Connelly, Chongle Pan, and Robert L. Hettich, Mass Spectrometric Approaches to Characterizing Bacterial Proteomes. *Expert Review in Proteomics* (2004), 1, 433-445.

The large amounts of information generated in the genomics era have begun to reveal the complexities of microbial systems. For example, complete genome sequence reveals the blueprint for life, in that it includes all information about the genes and gene products used by the organism for all of its life functions. This level of global genome information about an organism now makes it possible to begin to pursue an integrated approach to understanding how these organisms live and function *by cataloging and understanding all of the biological components, their functions, and all of their interactions in a living system and communities of living systems* [1]. A natural extension of *genomics* (the study of the complete set of genes for an organism) research is the characterization of the gene products, most of which are proteins. This latter research area is defined as *proteomics* (the study of the entire suite of proteins from a genome). Proteome analyses, whether in simple microbes, yeast, or higher organisms, present a much greater challenge than the genomics sequencing efforts. *While the genome is relatively static, the proteome is very dynamic*. The genome generally contains a set number of copies of every gene; however, proteins in the proteome can be expressed in a wide concentration range, varying from only a few copies per cell for regulatory proteins to many thousands per cell for ribosomal subunits.

Proteins are complex 3D structures, which constitute the machinery of a cell and at any time point perform the structural, catalytic, and signaling processes critical to cellular life. To aid in these complex processes, proteins often contain post translational modifications (PTMs); more than three hundred of these modifications have been identified to date [2].The term post translational modification (PTM) refers to modifications that occur during or after translation of the polypeptide chain.These post translational modifications are important to provide protein heterogeneity, thereby allowing a protein to exist in multiple isoforms. Most proteins must be modified in one or more of a number of ways with PTMs before they achieve their final functional form. PTM categories include: (a) covalent modifications such as phosphorylation, methylation, and glycosolation; (b) proteolytic processing e. g., the removal of signal and or pre-peptide sequences; (c) nonenzymatic modifications including deamidation and racemization. Some common modifications found in bacteria and therefore addressed in this study include: N-terminal methionine truncation, acetylation, methylation, phosphorylation, and the removal of signal sequences.

The first of these modifications is the N-terminal methionine truncation, in which the N-terminal residue of the newly-synthesized protein is modified in bacteria to remove the formyl group. The N-terminal methionine may also be removed by certain methionine aminopeptidases. The truncation of the N-terminal methionine depends on the charge and size of the amino acid side chain occupying the next position from the N-terminal methionine. The truncation event follows what is known as the “N-end rule”. This rule states that residues bearing small uncharged side chains, such as alanine, which are considered stable, allow docking of methionine peptidases that cleave the N-terminal

methionine[3]. Also, there are approximately 12 destabilizing residues, according to the "N-end rule", that mark the protein for degradation by ubiquitin ligase. Therefore, biologically the truncation may relate to the half-life of the protein..

In the case of acetylation, the amino-terminal residues of some proteins are acetylated, as well as lysines and arginines within the protein sequence. The biological significance of amino-terminal modification varies; some proteins require acetylation for function whereas others that are acetylated do not absolutely require the modification. It is possible that only a subset of proteins actually requires this modification for activity or stability, whereas the remainders are acetylated only because their termini fortuitously correspond to consensus sequences. Proteins with serine and alanine termini are the most frequently acetylated, and these residues, along with methionine, glycine, and threonine, account for over 95% of the amino-terminal acetylated residues [4, 5]. Only a subset of proteins with any of these amino-terminal residues are acetylated, however, none of them guarantees acetylation [6]. The complexity of the termini that are acetylated is due to the presence of multiple N-acetyltransferases (NATs), each acting on different groups of amino-acid sequences and whose specificity is determined by two or more residues at the amino-terminal positions [7]. Amino-terminal acetylation does not necessarily protect proteins from degradation, as has often been supposed, nor does it play any obvious role in protection of proteins from degradation by the 'N-end rule' pathway that determines whether to degrade proteins according to their amino-terminal residue.

The second common class of modifications includes amino acid side chain modifications. Common examples of these side chain modifications include methylation, acetylation, and phosphorylation. Methylation is an example of a common PTM found

primarily on lysine and arginine. These two residues have very polar side chains that are positively charged. When these residues are blocked by a methylation, the basic nature of that site within the protein can be changed, thereby making it more or less accessible to other protein targets. Also, when the basic nature of lysine and arginine are changed, it may serve to alter the protein structure. Many proteins have conformations that are pH dependent, and when altered unfold or fold in a new configuration; methylation may play a role in this process. Finally, within this class of side chain modifications is phosphorylation. Phosphorylation of proteins (at Ser, Thr, Tyr and His residues) is an important regulatory mechanism. For example, phosphorylation of tyrosine residues is an important aspect of signal transduction pathways, and bacterial cells sense and respond to environmental signals through histidine phosphorylation [8]. The final category is proteolytic processing, or the removal of signal and or pre-peptide sequences. As a protein is being synthesized, decisions must be made about sending it to the correct location in the cell, where it will be required. The information for doing this resides in the nascent protein sequence itself. Once the protein has reached its final destination, this information may be removed by proteolytic processing. This class of proteins all contains an N-terminus termed a signal sequence or signal peptide. The signal peptide is usually 13-36 predominantly hydrophobic residues, flanked on the N-terminal side by one or more positively charged amino acids such as lysine or arginine, and containing neutral amino acids with short side-chains (such as glycine or alanine) at the cleavage site. The signal peptide is recognized by a multi-protein complex termed the signal recognition particle (SRP). As proteins with signal sequences are synthesized, they are bound by the SecB protein. This prevents the protein from folding. SecB delivers the protein to the cell

membrane where it is secreted through a pore formed by the SecE and SecY proteins. Secretion is driven by the SecA ATPase. After the protein has been secreted, the signal sequence is removed by a membrane bound leader peptidase [9].

Understanding these complex PTMs is often a difficult task. However, difficulties exist, progress has been made toward identifying PTMs across multiple microbial species. One of the major goals of this dissertation was to develop methods for the identification of PTMs from microbial species under multiple growth conditions (Chapter 7). The two chosen species include *Rhodopseudomonas palustris* and *Escherichia coli*. *Rhodopseudomonas palustris* belongs to the α- proteobacteria and is a purple nonsulfur anoxygenic phototrophic bacterium found in diverse environments from fresh water to soil. One of the unique features of *R. palustris* is its ability to grow and function under many metabolic states. These states include: photoheterotrophic, where energy is obtained from light and carbon from organic carbon sources; photoautotrophic, where energy is from light and the main source of carbon is from carbon dioxide; chemoheterotrophic, in this state carbon and energy are from organic compounds; and finally chemoautotrophic, where energy is from inorganic compounds and carbon from carbon dioxide [10, 11, 12]. These multiple growth states provide the wild type *R. palustris* (strain CGA0010) with the ability to be a biofuel producer by generating hydrogen gas as a byproduct of nitrogen fixation, as well as a greenhouse gas sink by converting carbon dioxide into cell mass.

Since most of these metabolic states can easily be attained in laboratory settings, *R. palustris* is an ideal model system for the study of diverse metabolic modes and their control within a single organism. Recently, the genome of *R. palustris* has been

sequenced, revealing a 5.4 Mb genome with 4836 potential protein encoding regions [13]. This sequencing and annotation effort, along with proteome profiling [121], protein-protein interaction studies, global gene knockouts [14], and transcriptome profiling [15] will provide a detailed systems biology characterization of this microbe.

The second microbe chosen for study was *Escherichia coli*. This microbe is a γ-proteobacteria and found commonly as a facultative anaerobe that colonizes the lower gut of animals but also survives when released into the environment. *E. coli* are rod-shaped bacteria that possess adhesive fimbriae. *Escherichia coli* has become a model organism for studying many of life's essential processes, partly due to its rapid growth rate and simple nutritional requirements. Researchers have well established information about *E. coli*'s genetics; and have completed many of its strains genome sequences. *E. coli* K-12, was the earliest organism to be "suggested as a candidate for whole genome sequencing" [16]. Several strains of *E. coli* have been sequenced and studied in detail. It has a single circular chromosome with 4,639,221 base pairs and 4288 protein-coding genes. Of these protein-coding genes, 38% have no attributed function. *E. coli* K-12's genome, has a 50.8% G+C content. Genes that code for proteins account for 87.8% of the genome, stable RNA-encoding genes make up 0.8%, 0.7% is made of noncoding repeats, and about 11% is for regulatory and other functions [16]. An interesting feature of *E. coli* K12 is the ability to develop antibiotic resistance to streptomycin through point mutations within the ribosomal proteins, and is the reason why this organism was used for study in this dissertation (chapter 4).

Characterization of a bacterial proteome typically refers to the comprehensive detection and identification of the entire suite of proteins expressed by the microbial cell.

The entire suite of proteins may not be expressed under one growth condition or time point, therefore multiple growth states or time points may be examined to look at the entire complement of proteins in an organism. One of the techniques of choice to perform these complex characterizations of proteins from within the cell, is mass spectrometry. Mass spectrometry provides a powerful method to measure ions of intact and fragmented molecules in order to provide molecular mass information, as well as ion manipulation capabilities for obtaining detailed structural information at the isomeric level, including differentiation of isomers in many cases. Originally, mass spectrometry was known for its use in small molecule evaluation, but advances in the 1980's made it possible to extend its applications to large biomolecules such as proteins, nucleic acids, and their complexes. These key advances included the ability to ionize these large molecules using two new techniques. The new ionization techniques of electrospray ionization (ESI) [17] and matrix-assisted laser desorption/ionization (MALDI) [18,19] provided a new way of forming gas-phase ions from these larger molecules. These advances enabled mass spectrometry to become a leading technology for proteome measurements, due to its inherent ability to identify proteins, including hypothetical species, at high mass accuracy, resolution, and throughput, even from complex mixtures [20,21].

Currently, there are two major methods for analyzing proteins by mass spectrometry. The *top-down* method involves measuring intact proteins, either with or without MS/MS of these intact proteins. This method was first introduced with electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry, ESI-FTICR-MS [22, 23, 24] and expanded to ion traps with novel ion-ion reactions [25]. In the *bottom-up,* method, intact proteins are digested with a protease such as trypsin,

Glu-C or cyanogen bromide (CNBr), and the resulting peptide mixtures are analyzed by MS or MS/MS. It should be noted that in this definition it does not matter whether the initial separations are performed on intact proteins or peptides; rather, the experiment type is defined by the species measured by the MS. Thus, 2D-PAGE of intact proteins followed by in-gel digestion and MS analysis is considered a bottom-up approach. The actual development of the bottom-up methodology cannot be traced to a single lab, but rather evolved from multiple labs using very different techniques including gel-based[26, 27, 28, 29, 30] and solution-based separations[31, 32, 33] followed by MS or MS/MS for protein identifications. These two general approaches can be summarized as follows:

Bottom-up proteomics: Protein mixtures (from cell lysate or protein complexes) are proteolytically digested (usually with trypsin), and the resulting peptide mixture is examined by mass spectrometry. The MS data are used to query a peptide database from the specific organism to identify the protein components of the original mixture. *This method is excellent for determining protein identities, but provides very limited information about the molecular form of the intact proteins.*

Top-down proteomics: Complex protein mixtures from cell lysates or protein complexes are examined directly by on-line or off-line MS. No digest is conducted; rather the intact proteins are measured with MS and MS/MS [34]. *This method provides fewer protein identities, but does give detailed information about the intact molecular forms of the proteins, including post-translational processing (small molecule additions, truncation, mutations, and signal peptides).*

Both techniques have advantages and disadvantages and will be discussed in detail below. Bottom-up proteomics is by far the more widely used method, mainly

because it is much simpler to conduct and does not require high performance MS instrumentation. The progress in the field of bottom-up proteomics has been staggering. It has now become possible (if not routine) to measure ~1000-1500 proteins from a microbe under a given growth condition with a high degree of confidence in a 1-3 day period, depending on the technology used. Furthermore, if enough mass spectrometers are assembled, this analysis can be rapidly repeated for protein identification for an organism under a variety of different growth conditions.

Bottom-up proteomics has become almost routine to perform, although, top-down proteomics has moved at a slower pace. This lag in development is primarily due to the following factors: liquid-based separations of intact proteins are more difficult than peptides, MS and MS/MS analyses of intact proteins are more difficult to conduct and interpret than peptides, the high performance MS instruments capable of adequate analysis of intact proteins from complex mixtures are fairly expensive and have not been designed for routine operation in most cases, and the algorithms to analyze MS/MS of intact proteins are not as well developed or commercially available. Even with these experimental challenges, top-down proteomics provides a level of information that the bottom-up technique does not, which is the *intact state of the protein*. Information on the intact state of the protein is critical, since proteins function as intact molecular species, not as a combination of simple, small peptides. Thus, a full understanding of the intact state of proteins (PTMs, truncation, mutations, and signal peptides) is necessary.

Bottom-up MS proteomics has become very powerful over the past five years, although, it is clear that this is an *indirect* protein identification technique, as the *intact* protein species are never measured directly, but rather only a fraction of the proteolytic

peptides for any given protein are identified. This leads to some concern that subtle aspects of the protein, such as the presence of isoforms, or post-translational modifications, might be missed by the bottom-up approach. This need for intact protein measurements in complex mixtures has prompted investigation into developing MS technologies for this task. At initial thought, this may seem straightforward based on the extensive past work on characterizing purified protein samples, in fact, this approach turns out to be a formidable analytical challenge for proteomes due to at least three factors. First, the protein molecules masses can range from 5–200 kDa, requiring high performance MS technology for accurate measurements. Second, the extreme heterogeneity of protein sequences gives rise to a substantial ionization suppression effect when very complex mixtures of proteins are examined. Thus, the proteins with the largest amount of surface charge will ionize most easily and will be over-represented in the mass spectrum relative to their abundance in the sample. This factor suggests that some type of pre-fractionation, or on-line chromatography, will most likely need to be used for intact protein measurements. Third, the unambiguous identification of larger proteins is difficult, due to the isotopic packet that confounds accurate mass measurements and the inability to extensively fragment these proteins, under tandem mass spectrometry conditions, to get complete sequence information. All three of these factors are much easier for peptides because of their lower molecular masses and more extensive fragmentation. However, research is underway in several laboratories and has shown remarkable progress in overcoming these challenges for the top-down approach. One particular factor that must be noted is that most of the developments of the top-down approach have focused on the experimental LC and MS measurement technologies. As a

result, the bioinformatics component is much less developed for the top-down data analysis.

One of the challenges in separating complex protein mixtures is keeping the proteins intact and soluble during the preparation/fractionation process. Because MS measurements do not require the proteins to be in their active forms, it is sometimes desirable to denature the entire complex mixture as early in the clean-up process as possible. While this usually inactivates cellular proteases, it also can cause undesirable protein precipitation in the samples. For the bottom-up MS approach, it is advantageous to denature and digest the complex protein samples as early as possible in the clean-up process. Because only peptides are measured, protein stability is not an issue for this method. In contrast, protein stability is critical for the top-down MS approach. To enhance this, during the cellular lysing process, a protease inhibitor cocktail is often added to arrest protein degradation. The protease inhibitors, which are often small molecules, stabilize the protein samples, but can often be removed prior to MS characterization.

The critical component for top-down proteomics by MS is measurement of the molecular masses of the intact proteins. The five important experimental aspects of this measurement are *mass accuracy*, *mass resolution*, *dynamic range*, *mass range*, and *detection sensitivity*.

(i) *Mass resolution*. The measure of how well adjacent peaks can be differentiated in the mass spectrum. This value is typically given as the peak full width at half maximum (FWHM).

(ii) *Mass accuracy*. The comparison of the measured mass to the calculated mass. This value is typically given as error in either percentage or parts-per-million (ppm).

(iii) *Mass range*. The difference between the largest and smallest molecular mass that can be measured.

(iv) *Detection limits*. The smallest amount of sample that can be measured with a signal/noise of at least 3:1.

(v) *Dynamic range.* The molar difference between the least abundant component and the most abundant component that can be detected in a single sample.

The wide molecular range of possible proteins experienced in proteomics has researchers proposing the use of technologies such as MALDI-TOF-MS. This approach does provide an advantage for the analysis of large protein species, but does have some draw backs such as limited mass resolution and accuracy. For example, a protein with a molecular mass of 50 kDa can generally only be measured using a TOF-MS to about 0.02% (~ 10 Da). While this mass measurement is far superior to what is obtainable from gel electrophoresis, this value could still correspond to many proteins within a given database. Therefore, a much more accurate measurement, providing a higher level of mass accuracy, is needed to limit the number of possible proteins identifications from employed databases. This is the driving force to employ techniques such as ESI-FTICR-MS for intact protein measurements. This technology provides unprecedented capabilities for high performance measurements, although, many experimental parameters are difficult to employ and need further development. For example, the same protein with a molecular mass of 50 kDa could be measured with the FTICR-MS

technique to about 0.0005% or 5 parts-per-million (~0.25 Da). *Thus, high resolution and accurate mass measurements of intact proteins are often sufficient information to identify many bacterial proteins, without further structural information.* However, this statement is true in many cases, confounding the identification of intact proteins are protein truncations and post-translational modifications that alter the measured molecular masses, making it difficult to correlate the measured protein mass with the value predicted from the genome data. For this reason, it is best to integrate the measured molecular mass information with either structural data obtained by tandem mass spectrometry or with data obtained by the bottom-up MS method on the same organism [35].

High-resolution molecular mass measurements of intact proteins reveal the complex isotopic packet resulting from the combination of naturally-occurring isotopes. This necessitates comparing the measured and calculated isotopic distributions to verify protein identification [36]. In practicality, the high-resolution molecular mass measurement is used to query a protein database for a given organism. The possible protein matches falling within the specified mass accuracy window are tabulated, and a calculated isotopic distribution is determined for each one (for FTICR-MS measurements, there are usually no more than 3-4 possible proteins within the 5-10 ppm range of the measured mass). For each putative protein, the calculated isotopic distribution and most abundant peaks are compared to the measured values for final protein identification.

Even with the high-resolution molecular mass measurements discussed above, the dynamic range and heterogeneity of intact proteins in these complex mixtures can confound the MS measurements. The basic problem stems from the limited ability to

simultaneously measure hundreds (or even thousands) of proteins in a single mixture. An obvious solution to this dilemma is to incorporate some aspect of protein fractionation, either off-line or on-line, with the MS measurement. This increases the sample handling and possible contamination or sample losses, but the MS measurement requirements are greatly relaxed. For example, off-line anion-exchange chromatography can be used to fractionate complex protein mixtures from crude cell lysates. Each fraction, which contains between 50-200 proteins, is more easily interrogated by mass spectrometry [35].

The most common protein fractionation approach has been to incorporate reverse-phase liquid chromatography on-line with the MS. This arrangement permits the proteins to be physically separated by their hydrophobicity on the stationary phase of the column, and then eluted, sequentially, directly into the mass spectrometer. Reverse phase chromatography columns, employed in this research, have a stationary phase composed of silicate which has reactive hydroxyl groups. In order to cap these hydroxyl groups and keep them from reacting with the proteins, alkyl chains are added. The longer the alkyl chain caps of the silicate ends, the further the proteins are from the reactive hydroxyl groups. Generally, most peptide work employs a C18 stationary phase for the best separations. However, this is not the case for intact proteins wherein the shorter the carbon backbones within the stationary phase generally mean better separation of intact proteins. This need for shorter carbon chains is due to the large size and variation of hydrophobicities of intact proteins. Therefore, most intact protein separations employing reverse phase chromatography use a C2 to a C4 carbon backbone. This form of separation and measurement takes longer, (usually about 2 hours for the LC-MS experiment), although, a much more extensive analysis of the complex protein mixture is

possible. This approach has been demonstrated for the characterization of the chloroplast grana proteome [37] and the yeast large ribosomal subunit [38], and resulted in not only protein identifications but also detection of post-translational modified species. It is feasible to employ a multi-dimensional chromatographic approach for more enhanced protein fractionation. For example, a two-dimensional LC-MS experiment has been conducted on *Saccharomyces cerevisiae* by using a version of gel electrophoresis employing acid-labile surfactants, followed by reverse-phase LC directly into an FTICR-MS [39].

There are several alternatives to on-line chromatography. One such approach involves surface enhanced laser desorption/ionization TOF-MS approach [40]. For this method, a variety of chemical (hydrophobic, ionic, or mixed) or biochemical (antibody, DNA, enzyme, or receptor) surfaces are used to preferentially absorb selected protein species. This allows the fractionation to be fairly generic or highly specific, thereby selectively reducing the complexity of the protein sample. These surfaces can be incorporated into protein chips, providing a high-throughput sampling methodology for MALDI-TOF-MS, although the identification of proteins from only their low-resolution molecular mass is difficult. Another alternative to liquid chromatography focused on exploiting the demonstrated power of gel electrophoresis. As a modification of conventional 2-D PAGE, mass spectrometry has been used to replace the size-based separation component of the SDS-PAGE separation [41]. For this method, the proteins separated according to pI are then measured by MALDI-TOF-MS, with either post-source decay dissociation of intact proteins, or peptide mass mapping experiments. Such information can be used to construct virtual 2-D gels.

To unambiguously verify the protein assignment by top-down MS, it is advantageous to acquire at least some structural information for the intact proteins [23,42]. This can be accomplished with a variety of tandem mass spectrometry experiments, involving collisional dissociation, electron dissociation, or photodissociation. Proteins usually fragment much less extensively than peptides, but there is often sufficient fragment ion information to confirm or reject a possible protein identification from the accurate mass measurement. For example, the presence of only three or four fragment ions from a protein was found to be sufficient for a 99.8% probability of identifying the correct protein from a database of 5,000 bacterial protein forms [43]. This methodology can be applied for proteins both with and without disulfide bonds [44,45]. Electron capture dissociation shows promise for the most extensive fragmentation of intact proteins in a high-throughput manner [46,47]. Electron capture dissociation uses low-energy electrons to neutralize the charges on the protein producing cleavage of the amide bond to form c and z ions, and usually provides extensive sequence coverage of proteins even up to 45 kDa in size [48]. A combination of collisional dissociation and electron capture dissociation can be used to provide complementary information on intact proteins in bacterial proteomes [49]. Collisional activated dissociation (CAD) traditionally has been one of the most common fragmentation methods for proteins in top-down mass spectrometry. CAD is capable of producing high fragmentation efficiency with relatively simple implementation [50]. For very large proteins (molecular masses exceeding 150 kDa), it may be advantageous to employ partial proteolytic digestion to make large peptides (5-50 kDa), and then characterize these species [51]. One of the more extensive techniques for top-down MS

is a combination of capillary LC-MS with infrared multiphoton dissociation (IRMPD) [52]. IRMPD offers a method of fragmentation where no single frequency excitation is required and the ions of all m/z values are dissociated at the same time [53]. This method has been demonstrated to be useful with proteins and peptides.

As stated above, the bottom-up and top-down MS approaches each have unique capabilities and limitations. One approach to exploit the power of each technique is to integrate them together, with the goal of more comprehensive proteome characterization. A flow-chart describing how this integrated technique might be designed is illustrated in Figure 1.1. Off line fast protein liquid chromatography (FPLC) is used in this integrated method to separate the large complex mixtures of proteins mixtures for top-down analysis due to its proven ability to reduce down the complexity of the mixture. Therefore, by reducing the complexity of the protein mixture, this method allows for better separation from the on-line HPLC methods used, as well as more comprehensive protein identifications [35]. This method of off line FPLC fractionation followed by on line HPLC does take a large amount of protein starting material this is not of great concern due to the ability to produce more than enough material from the chosen microbe's cultures. Another area of concern using this strategy is the loss of protein during the off line separation. This problem is unavoidable due to the need to have a prior separation of the complex protein mixture before the top-down analysis.

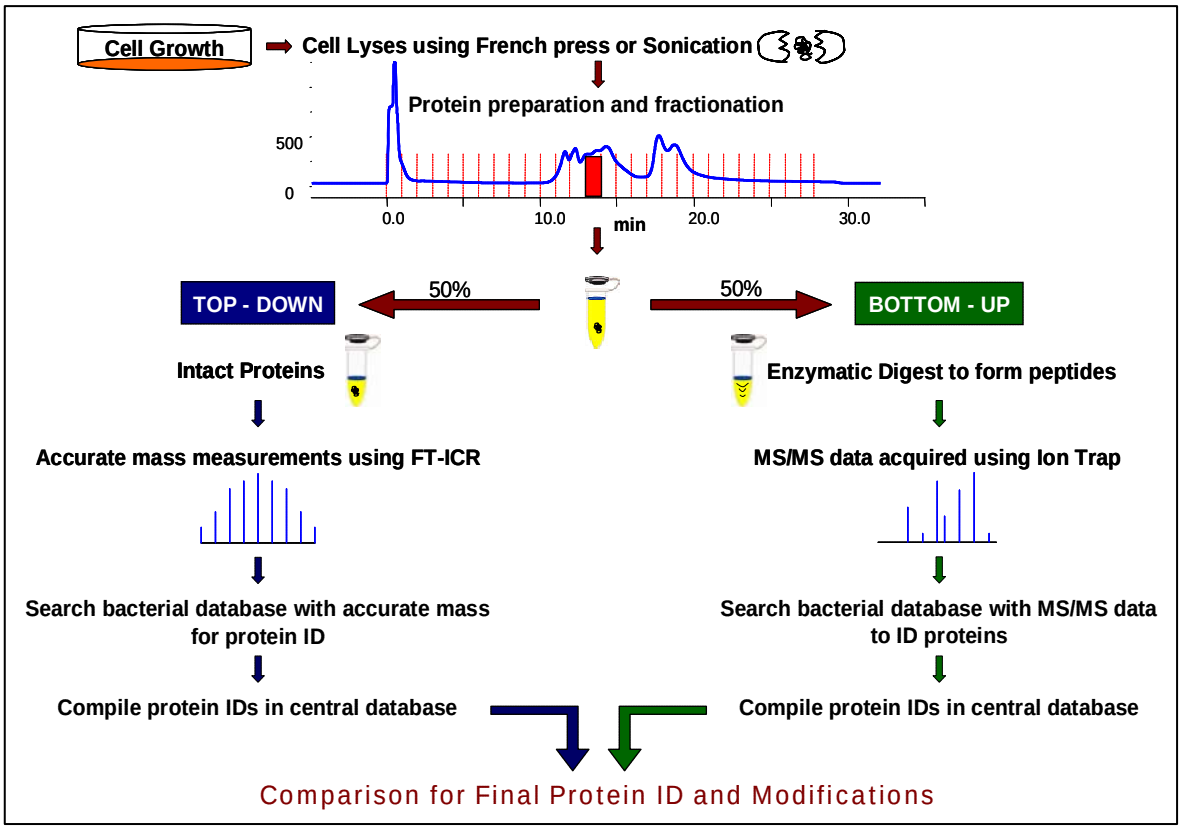


Figure 1.1: Integrated protein preparation and identification.
Flow chart illustrating how an integrated top-down and bottom-up MS approach can be used to characterize a bacterial proteome.

By providing this initial separation step with FPLC, we increase the overall ability to analyze and identify more proteins than with no initial separation. We have found this technique of off line fractionation followed by on line HPLC to be highly reproducible and simple to implement for a large-scale study of multiple samples [35]. An integrated top-down and bottom-up approach allows for a more complete characterization of protein complexes due to the unique strength of each technique. In an integrated approach, intact protein masses, from the top-down analysis, corresponding to a particular PTM or isoform, are then able to be compared to the comprehensive list of proteins provided by the bottom-up analysis. This correlation between the two methods can provide PTM location and identity with more certainty. The comprehensiveness of this technique has been previously demonstrated in studies of the *Shewanella oneidensis* [35] proteome as well as the 70S ribosomal complex from *Rhodopseudomonas palustris* [54].

The major goal of this dissertation was to build a platform for the analysis of intact proteins from complex mixtures, in order to obtain information about the natural state of the proteins. The hope was to gain greater biological insight into the complex systems of microbes by providing starting information about the function, and possible cellular location of proteins from bacteria. At the start of this dissertation, top-down proteomics was only beginning to be developed in numerous laboratories. Thus, a major effort was needed to develop the necessary biological, analytical, and computational tools to addresses this daunting technical challenge of analysis intact proteins. The research presented here has helped to bring us one step closer to achieving that goal.

The following is an outline of that effort. Chapter 2 details the current ORNL

"top-down" proteomics pipeline for microbial proteomics, which was developed primarily through efforts of this dissertation. Chapter 3 details the fundamental work on the FT-ICR for the evaluation of proteins and PTMs. These fundamental efforts were needed to advance this dissertation work on proteins and PTMs. Chapter 4 illustrates our evaluation of complex ribosomal mixtures for PTMs and isoforms from the two microbes *R. palustris* and *E. coli*. Chapter 5 further illustrates the effectiveness of examining PTMs in protein complexes for key regulation sites from *Rhodopseudomonas palustris*. Chapter 6 introduces new computational methods developed for integrated top-down and bottom-up data for the identification of PTMs. Finally, Chapter 7 concludes with the application of "top-down" proteomics for the first characterization of a microbial proteome from multiple growth conditions. This dissertation is the culmination of years of effort to develop a top-down proteomics platform for the characterization of intact proteins and PTMs from microbial proteomes with differing environmental conditions.

Chapter 2

Experimental Platform for the Analysis of Intact Proteins and PTMs in Microbial Systems by Mass Spectrometry

Introduction

This chapter describes the experimental platform for analysis of intact proteins and their associated post translational modifications (PTMs) from either protein complexes or microbial cell extracts that was developed through the course of this dissertation. While a common experimental thread of analyzing intact bacterial proteins for PTMs and isoforms by an integrated top-down and bottom-up mass spectrometry approach can be found in all following chapters, the exact methods vary to some degree. This chapter breaks each part of the process down and explains variations and advantages and disadvantages of the various methods. The ORNL integrated top-down and bottom-up platform is illustrated in Figure 2.1. The major parts include cell growth, protein extraction/sample preparation, liquid chromatography, mass spectrometry, proteome informatics and biological information extraction. Each of these subtasks are detailed below.

Cell Growth and Protein Preparation

For all studies presented in this dissertation bacteria were grown from stock solutions in batch format. Generally, glycerol stock solutions of the WT strain or a mutant strain are kept at -80^{o}C. For the *R. palustris* studies in Chapters 4, 5 and 7, the wild-type (WT) strain CGA0010 was a gift from Dr. Caroline Harwood at the University of Washington and can be obtained from Dr. Dale Pelletier in the Life Science

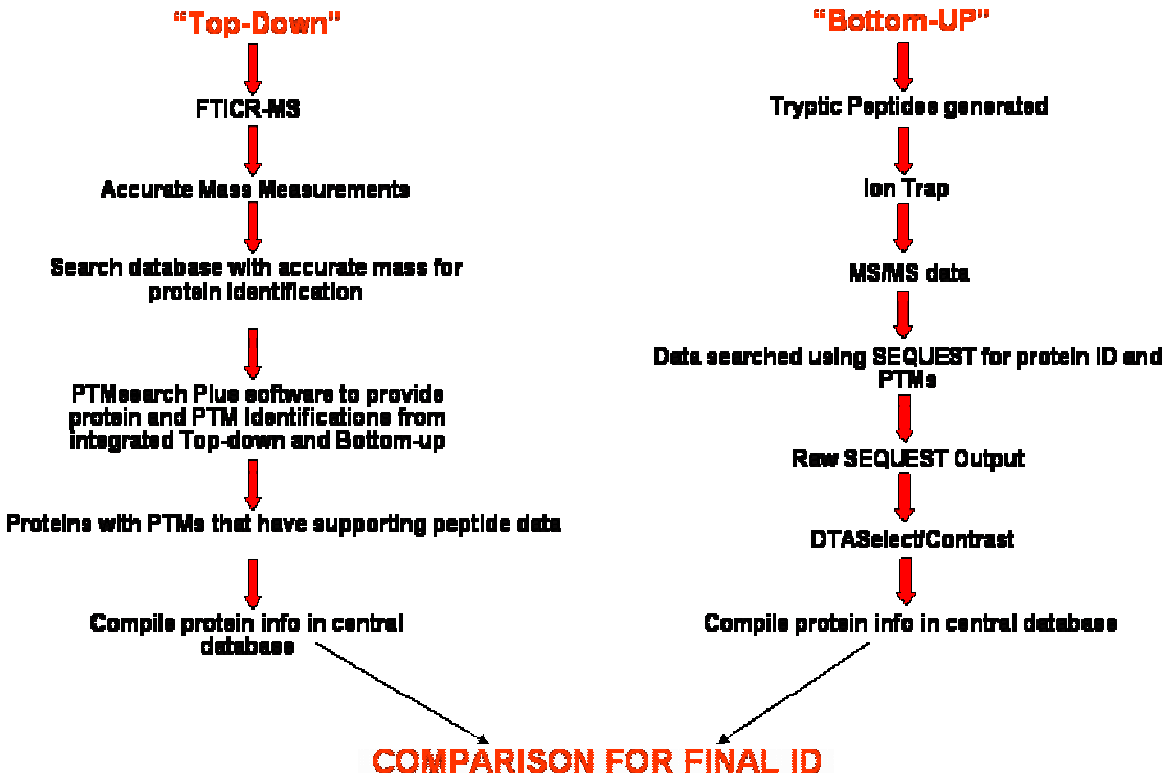


Figure 2.1: Major steps in integrated top-down and bottom-up proteomics pipeline. Illustrated is each major step in the ORNL proteomics pipeline for the analysis of individual protein complexes and entire proteomes.

Division at ORNL. The *Escherichia coli* protein purifications that were used in the antibiotic resistance work in Chapter 6 were supplied by Dr. Morgan Giddings at the University of North Carolina already in the purified intact protein form.

Growth of Wild Type R. palustris

R. palustris strain CGA0010, a hydrogen-utilizing derivative of the sequenced strain (unpublished C. S. Harwood) and referred to here as the wild-type strain, was grown under the three conditions (chapter 7). Wild type *R. palustris* cells were grown anaerobically in light or aerobically in dark on defined mineral medium at 30 °C to mid-log phase (OD660nm = 0.6). Carbon sources were added to a final concentration of 10 mM succinate and 10 mM sodium bicarbonate. For the photoheterotrophic N_2 fixing cultures, ammonium sulfate was replaced by sodium sulfate in the culture medium and N_2 gas was supplied in the headspace. Chemoheterotrophic cells were grown aerobically in the dark with shaking at 200 rpm; phototrophic cells were grown anaerobically in the light with mixing with a stir bar. All anaerobic cultures were illuminated with 40 or 60 W incandescent light bulbs from multiple directions. 4-5 liters of cells were grown for all three states and pooled together for each state.

Protein Extraction of Wild type R. palustris

The cell pellet from each growth state were resuspended in ammonim acetate buffer then lysed using a French Press. Total protein yields range between 60-120 mg of protein for each of the three growth states. Cell extract was centrifuged at 10,000 X g for 35 minutes in a Sorvall centrifuge to remove all unbroken cells. Protein extract was used for off-line anion exchange FPLC fractionation. Anion Exchange fractionation was used due to the pI range of most proteins is in 3-7 range. By employing anion exchange with

buffers in the pH range of 7.5-8, most proteins will not reach their isoelectric point and will be eluted off the column according to their pI. Illustrated in Figure 2.2 is the protein isolation process followed by mass spectrometry. To perform off-line anion exchange chromatography 60 mg of protein was injected onto a 5 ml HiTrap (HiTrap SP HP, Amersham Pharmacia) ion exchange column connected to an AKTA (Amersham Pharmacia) FPLC system. After protein injection a 30 minute ammonium acetate gradient was run from 0.2 M to 2 M at pH 7.5. Twenty fractions from each growth state (total of 60 from 3 growth states) were determined to have sufficient protein amounts (400 μg) by a Bradford protein assay. Each FPLC fraction obtained was then divided into two portions. One portion was examined by 1D LC-MS-MS bottom-up mass spectrometry and the other portion of the sample was examined using LC-FTICR-MS for top-down mass spectrometry.

Creation of Affinity Tagged Proteins in R. palustris

The *R. palustris* wild type strain (CGA0010), harboring the pBBR5-DEST/42 modified Gateway expression plasmid (Invitrogen, Carlsbad, CA) with the RPA0274, RPA0272, RPA2966 open reading frames (ORF) were generated at Oak Ridge National Laboratory by Dr. Dale Pelletier. The ORFs were cloned into the expression plasmid with the V5 and 6xHis affinity tags fused at the C-terminus of the protein.

R. palustris cells harboring the expression plasmid were grown anaerobically and under nitrogen fixing conditions in PM-N2 (photosynthetic nitrogen fixing medium) or in PM (photosynthetic medium) under non-nitrogen fixing conditions. Cells were harvested at mid-log phase ($O.D._{660}$ ~0.8). Cell pellets were re-suspended in NTA binding buffer (50 mM NaH_2PO_4 at pH 8, 300 mM NaCl, 10 mM imidazole, 5 mM ATP, and 10 mM

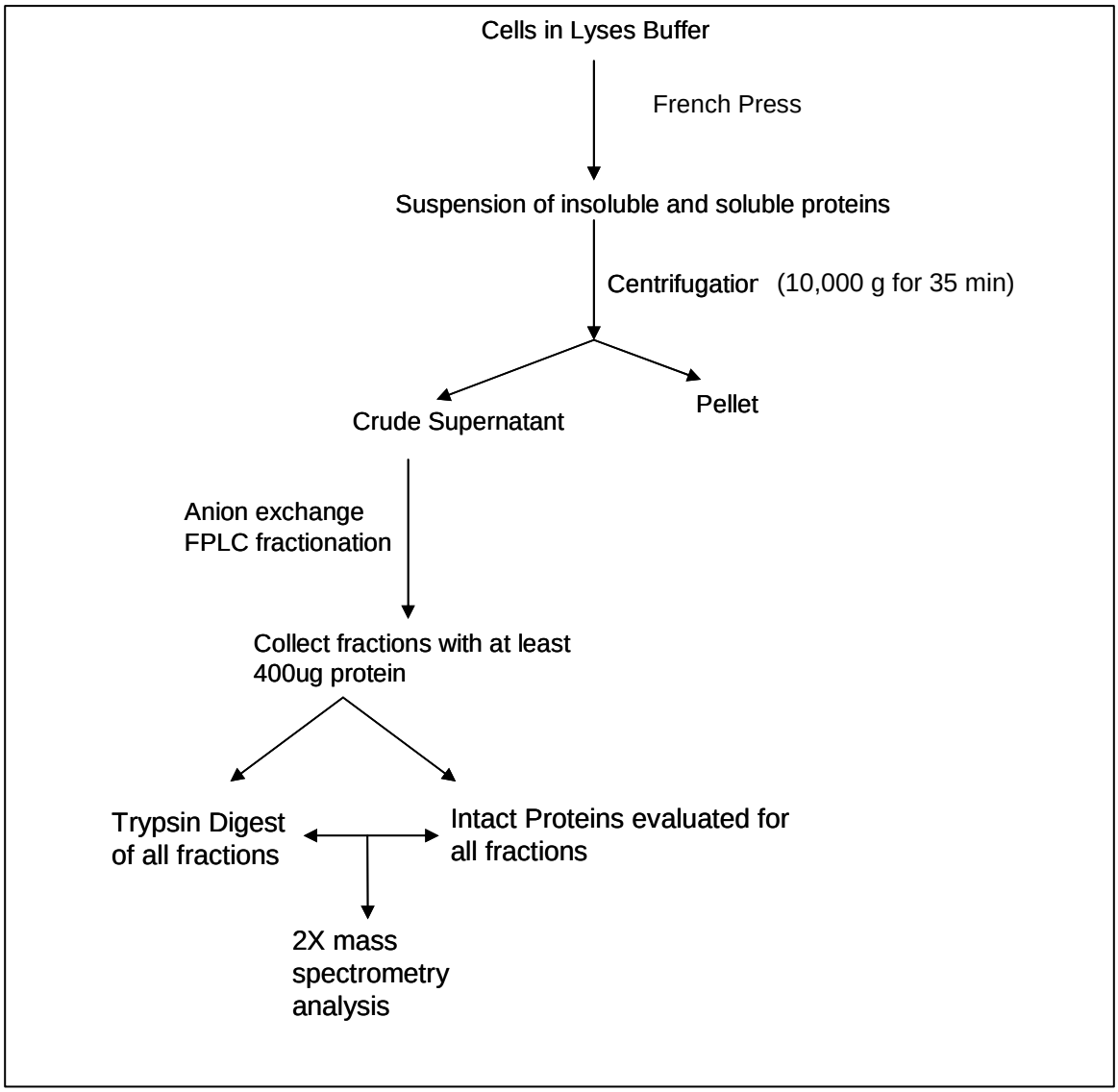


Figure 2.2: Steps in protein purification performed.
Illustrated are the steps in cell growth to protein purification employed in dissertation.

$MgCl_2$, 10 mM KCl, 100 µg/ml PMSF and 10 µg/ml leupeptin) and lysed with 1X BugBuster (Novagen). Cellular debris was removed with an initial centrifugation at 4°C using an SS-34 Sorval rotor at 12,100 × *g* for 30 minutes. The supernatant was centrifuged for an additional 15 minutes at 23,700 × *g*. The final resulting supernatant was then immediately used in the first stage of the affinity purification.

Affinity Purification

The presence of two tags (6X His-tag and V5 antibody tag) within the expressed protein allowed for the use of a dual affinity purification strategy to "capture" the complexes. Figure 2.3 illustrates the affinity purification process employed. This is a standard protocol for large-scale isolation of protein complexes from *R. palustris* in our laboratory, in which a large number of strains each bearing a plasmid encoding a different affinity-tagged protein [55].

In the first purification step, Ni-NTA beads (Qaigen, Valencia, CA) (previously washed in NTA Binding buffer 4X) are added to the supernatants and were incubated on a rotator for one hour at ambient temperature. The beads were then collected by centrifugation at 425 × *g*, transferred to new tubes, and washed 4X with NTA wash buffer (50 mM NaH_2PO_4 at pH 8, 300 mM NaCl, 20 mM imidazole, 5 mM ATP, 10 mM $MgCl_2$, 10 mM KCl). Afterwards, bound proteins were eluted from the Ni-NTA beads 4X with NTA elution buffer (50 mM NaH_2PO_4 at pH 8, 300 mM NaCl, 500 mM imidazole, 5 mM ATP, 10 mM $MgCl_2$, 10 mM KCl). Combined eluents (approximately 150 µl total) were diluted with 400 µl buffer (5 mM ATP, 10 mM $MgCl_2$, 10 mM KCl) and immediately used for the second affinity purification step.

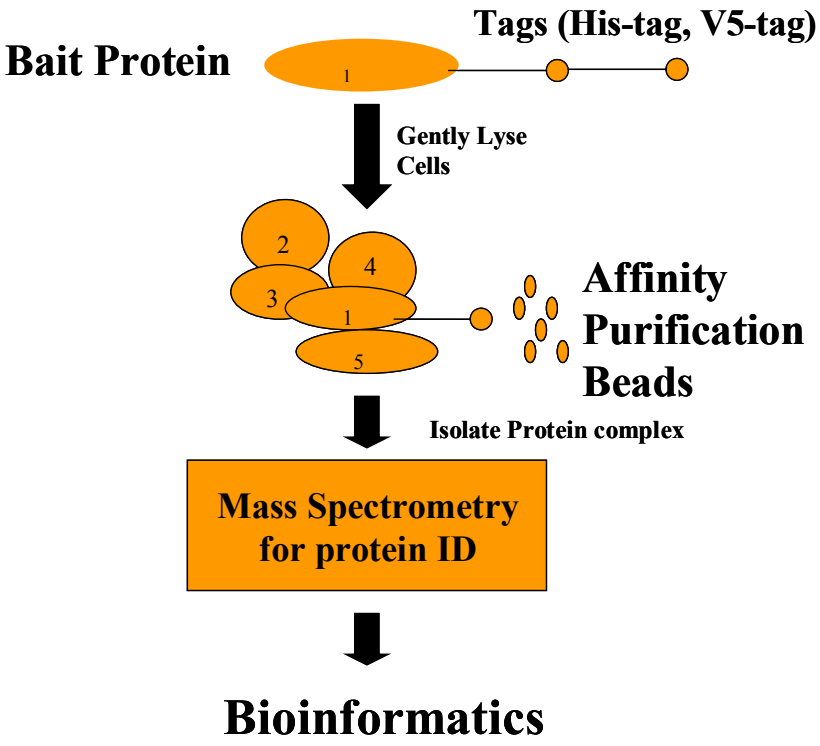


Figure 2.3: Steps in protein affinity purification performed.
Illustrated are the steps in affinity protein purification employed in dissertation.

V5 beads (Sigma, St. Louis, MO) (previously washed in PBS buffer) were added to the combined eluents from the Ni-NTA capture and incubated on a rotator for one hour at ambient temperature. The beads were then centrifuged at 425 × *g* and washed 4X with V5 wash buffer (50 mM Tris-HCl, 10 mM $CaCl_2$ at pH 7.6, 5 mM ATP, 10 mM $MgCl_2$, and 10 mM KCl). Afterwards, the bound proteins were eluted three times from the V5 beads with V5 elution buffer (80% acetonitrile and 1% formic acid). The combined eluents were analyzed by protein chip measurements to give total protein concentration of 5µg in 150µl of eluent. This affinity purification method was completed 4 times to provide 2 purifications for Top-down mass spectrometry analysis and two purifications for bottom-up mass spectrometry analysis of each *R. palustris* growth state. Approximately 10 µg of affinity purification eluent from each growth state was digested for bottom-up analysis with sequencing grade trypsin added at 1:20 (wt/wt) of protein to enzyme. The digestions were run with gentle shaking at 37 °C for 12 hours. Samples were immediately desalted with an Omics 100 µl solid phase extraction pipette tip (Varian, Palo Alto, CA). All samples were frozen at -80°C until LC-MS/MS analysis.

FTICR-MS

Rationale for Using Fourier Transform Ion Cyclotron Resonance Mass Spectrometry (FTICR-MS) for the Characterization of Intact Proteins and PTMs in Microbial Systems

The analysis of large bio-polymers i.e. proteins, and their associated complexes is a current area of scientific investigation addressed in this dissertation. Fourier Transform-Ion Cyclotron Resonance-Mass Spectrometry (FTICR-MS) is an analytical tool that has found particular application in the area of biological mass spectrometry [56]. FTICR-MS is particularly suited to the analysis of intact proteins as well as peptides [57]

because of its unique method of mass analysis and m/z determination. FTICR-MS provides mass resolution (FWHM of 100,000 to 150,000) far superior to other types of instruments and also provides high mass accuracy (1 to 10 ppm for molecules of 100 to 30,000 Da) with proper calibration [57, 58]. In addition, its ability to comprehensively measure a wide dynamic range (up to 10^5) provides an exceptional tool for the analysis of complex mixtures. FTICR-MS has mass resolving power unparalled by other mass analyzers consequently, what appears as an unresolvable mixture with other techniques appears as a data rich mass spectrum. This resolving power can be utilized at low mass as well as high mass applications. However, it is important to remember, the FTICR-MS resolving power does decrease with increasing mass to charge.

The high performance that can be achieved only by FTICR-MS was particularly crucial for analyzing intact proteins and their modified forms. Because of multiple carbon atoms in the molecule, the molecular region of the protein exists as a population of numerous isotopic species. The mass of a protein is determined most accurately if different isotopic species are resolved. Even for smaller proteins, FTICR-MS is only the instrument that can comprehensively resolve all of these isotopic species. Resolution of isotopic species is even more important when analyzing modified proteins. For example, as described in Chapter 4, the GlnK proteins in *R. palustris* are modified with an uridylylation. The modified forms of the protein have a mass shift of 306.2 Da and are difficult to resolvable in other low resolution instruments, such as ion traps. Another PTM that is sometimes difficult to resolve in lower resolution instruments is the methylation. The mass of methylation is 14 Da, which is very close to the mass of other common side chain losses such as water or ammonia (18 and 17 Da, respectively). Such

mass differences of intact proteins can be probed only when isotopic species are comprehensively resolved. In the next section, the fundamental principles of FTICR-MS will be illustrated. A more detailed description of FTICR-MS can be found in [57, 58].

Basic Principles of FTICR

In all FTICR-MS experiments preformed ions were generated in an electrospray source, de-solvated in a heated glass capillary, accumulated in a external hexapole, transferred into a high vacuum region with a quadrupole lens system, and then detected in the cylindrical analyzer cell of the mass spectrometer (Figure 2.4). Ion detection was achieved in an ultra low vacuum region (~2 X 10^{-10} Torr) through the use of differential pumping stages. Initial pumping was achieved using a mechanical pump which lowered the pressure to the millitorr range. The next stage of pumping was achieved using a turbo-pump to lower the pressure to ~ 10^{-5} Torr. Finally, two cryopumps lowered the base pressure to approximately 2 X 10^{-10} Torr. Once the ions reach the analyzer cell under the low pressure, the process of detection takes place. Detection in a FTICR is unique when compared to other mass spectrometers. FTICR-MS measurements rely on the cyclotron motion of ions in a magnetic field. This cyclotron motion is due to magnetic forces that bend the ion motion into a circle. The frequency of the ion cyclotron motion is unique to an ion of a particular mass/ charge. On the other hand, the frequency of the ion cyclotron motion is independent of ion velocity and proportional to magnetic field strength. Thus ions of a given mass to charge will have the same cyclotron frequency, regardless of the time the ion enters the cell or the velocity with which the ion enters the cell.

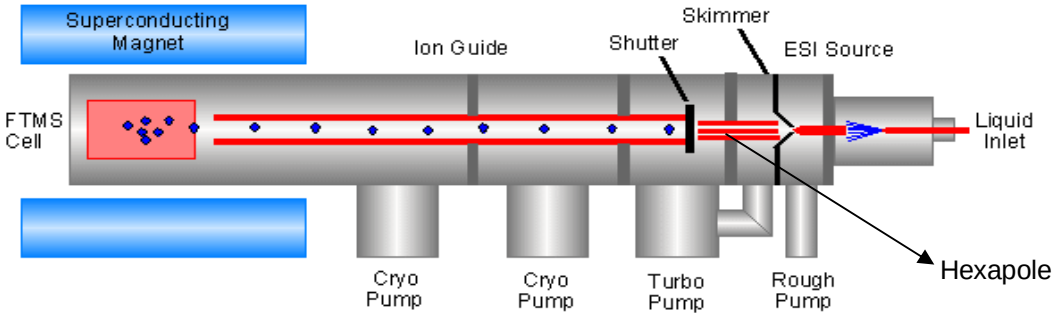


Figure 2.4: Schematic of IonSpec FTICR-MS.
Illustrated is the IonSpec ES-FTICR-MS instrument. Ions are introduced through the Analytica electrospray ion source and transferred through a heated glass capillary into a mechanically-pumped region, next through a skimmer, then into a turbopumped rf-only hexapole for accumulation and storage at 2 X 10^{-5} Torr. Finally, the ions are then gated through a shutter, down a quadrupole ion guide into the Penning cell within the high magnetic field. The penning cell is at ~10^{-10} Torr provided by two cryopumps. Figure is courtesy of IonSpec (www.IonSpec.com).

Measuring the cyclotron frequency permits ultra-high mass resolution [57, 58]. To put this into context, the base equation of ion cyclotron frequency can be examined [57].

$$\omega = q\ B/m$$

In this equation (ω) is the cyclotron frequency, (B) the magnetic field strength, (q) the charge of the ion and (m) the mass of the ion examined [57]. The frequency of the ion cyclotron is independent of ion velocity and proportional to magnetic field strength, and is inversely proportional to the mass/charge of the ion. In our case the magnetic field of the FTICR-MS used is 9.4T, therefore, the ion frequencies are in the radio frequency (rf) range of 10 kHz to 3 MHz. Using the above equation, the frequency of the ion cyclotron motion can be used to determine the mass/charge of an ion; what is ultimately measured in FTICR-MS.

It is important to understand that the ions are confined within the analyzer cell by an electrostatic potential and magnetic fields. The electrostatic potential is applied on two plates positioned perpendicularly to the magnetic field in the cell. Ions trapped in a magnetic field generally have incoherent cyclotron motion (i.e. they are moving independent of each other). In this mode, it is impossible to detect their net motion. To force the ions to move coherently, an electric field at the appropriate frequencies need to be applied. Normally the radius of an ion's orbit will be about 0.1 mm, but if an RF frequency is sent to the cell that is equal to the cyclotron frequency of the ion, it will gain energy from the *rf* field and move into a larger orbit. As a positively charged ion passes near the first electrode (forming part of the ICR cell), it will induce electrons toward the electrode. Then, as the ion moves away and approaches the second electrode, the electrons migrate to the second electrode instead [57].

While absorbing power, the ions are accelerated, and at the same time, all ions of the same mass/charge are forced to move in a phase coherent motion forming a packet of ions. This coherent ion cyclotron motion is called the ion cyclotron resonance (ICR). On the other hand, if the frequencies of *rf* and ion cyclotron are different, the ions will not absorb power. This time dependent migration of the electrons is converted to an image current by placing a resistor on the wire connecting the two electrodes, and the resulting image current is a sinusoidal signal (Figure 2.5). The signal produced is amplified and then fed into a computer. The amplitude of the image current is proportional to the number of ions within the ion packet. If there are only ions of a single mass/charge in the mass analyzer cell, the image current will resemble a pure sine wave. This sine wave can be expressed in the time domain as a function of voltage amplitude with respect to time. By a mathematical operation called Fourier transformation, the time image current can be converted into the frequency domain. In the frequency domain, amplitude is proportional to the abundance of ions trapped in the analyzer cell. In other words, the mass spectrum is a mirror image of the frequency domain. If there are ions with different mass/charge ratios, a complex waveform representing multiple image currents from the ion packets will be formed. In order to Fourier transform this waveform, it needs to be converted into a series of individual waveforms, called the Fourier series. In the Fourier series, the waveform is expressed as the sum of all the sine and cosine terms, therefore, forming the image current for an individual ion packet [57, 58]. The image current produced by the Fourier series is converted into the frequency domain, and the frequency domain is further converted into a mass spectrum (Figure 2.6).

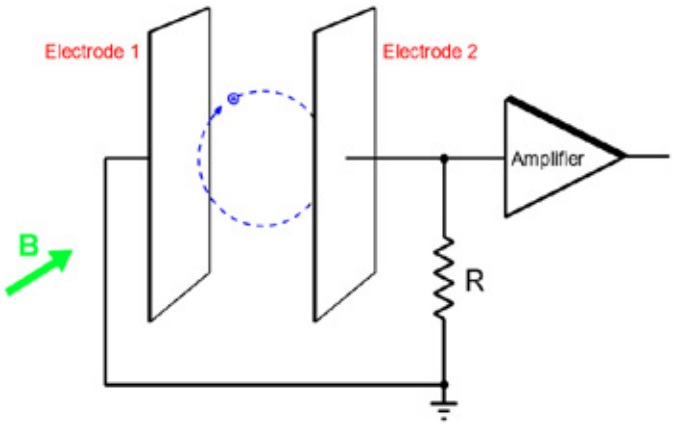


Figure 2.5: Generation of image current within the FTICR-MS.
(Image was taken from www.IonSpec.com)
Schematic of how the image current is obtained from the ion cyclotron frequency. The magnetic field is represented by the green (B) within the diagram.

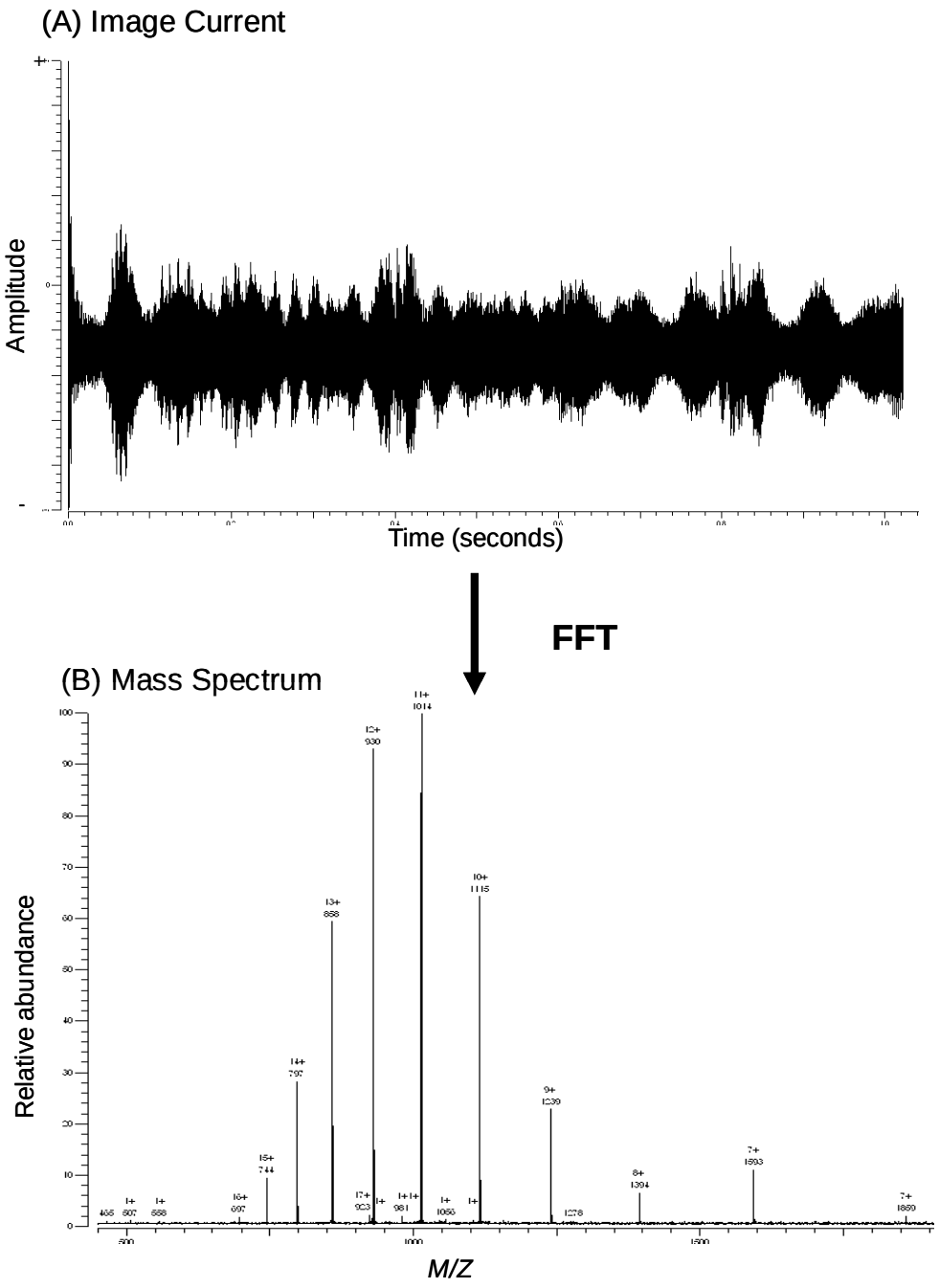


Figure 2.6: Generation of mass spectrum from the image current within the FTICR-MS. (A) The image current produced from a complex protein mixture. (B) The mass spectra obtained after a Fourier transform from the image current.

Experimental Procedure for Mass Spectrometric Analysis of Protein and Peptides Employed In This Study

ESI-FTICR Mass Spectrometry

All ESI-FTICR mass spectra were acquired with an IonSpec (Lake Forest, CA) 9.4-Tesla HiRes electrospray Fourier transform ion cyclotron resonance mass spectrometer. A Harvard syringe pump (flow rate of 1.75 μL/min) was used for direct infusion into an Analytica electrospray source (Analytica of Branford, CT). After generation, ions were accumulated in an external hexapole and transferred into the high-vacuum region with a quadrupole lens system. Detection then followed in the cylindrical analyzer cell of the mass spectrometer. Calibration of the mass spectrometer was accomplished externally with ubiquitin, resulting in a mass accuracy of ±3-5 ppm and mass resolutions of 50,000-160,000 (FWHM) as previously described [59].

Capillary HPLC-FT-ICR-MS

Capillary HPLC-FTICR-MS was accomplished with a Dionex UltiMate HPLC interfaced directly to the FTICR instrument. A C4 reverse-phase column (VYDAC 214MS5.325 C4 column 300μm id x 250mm, 300Å with 5μm particles, Grace-Vydac, Hesperia, CA) was employed for all separations. The *R. palustris* FPLC purification eluent and *E. coli* ribosome purifications consisting of 20-30 μg of total protein was injected onto the column and eluted at 4 μl/min into the electrospray ion source of the FTICR-MS. The gradient was run from 100% solvent A (95% water, 5% acetonitrile, 0.1% formic acid, 50 mM hexafluorisoproponal) to 100% solvent B (95% acetonitrile, 5% water, 0.1% formic acid, 50 mM hexafluorisoproponal.) over a 75-min. linear gradient. Hexafluorisoproponal (HFIP) was added as a chaotrope to help proteins unfold

and keep them from forming multimers, which gives better peak resolution. Ions were generated with a 3700 V potential between a grounded needle and heated transfer capillary. After generation, ions were accumulated in an external hexapole for two seconds and transferred into the high-vacuum region with a quadrupole lens system. Detection then followed in the cylindrical analyzer cell of the mass spectrometer. Calibration of the mass spectrometer was accomplished externally with ubiquitin resulting in a mass accuracy of ±3-10 ppm and resolutions of 50,000-160,000 (FWHM). Because the mass resolution was at least 50 000 for the intact protein measurements, the molecular masses of these proteins could be measured with isotopic resolution.

Quadrupole Ion Trap MS

The function of the quadrupole ion trap is as follows; preformed solution phase peptide ions are sprayed through an electrospray or nanospray source on the front of the instrument into a heated capillary. The heated capillary is generally set at 150-250^{0}C and functions to desolvate the ions. The ions are then directed through a tube lens and passed through a skimmer. The skimmer acts to focus the ion beam and skim off neutrals. Next, the ions are directed through a quadrupole and octopole, which acts as an ion beam guide to focus the ions into the ion trap. The ion beam enters into the trap through the inlet and is trapped through action of the three hyperbolic electrodes: the ring electrode, the entrance and exit the endcap electrodes (Figure 2.7) [60]. Various dc and rf voltages are applied to these electrodes which results in the formation of a potential well, in which ions are trapped. The ring electrode RF potential produces a 3D quadrupole potential field within the trap. This traps the ions in a stable oscillating trajectory within the trap to produce what is known as dynamic trapping.

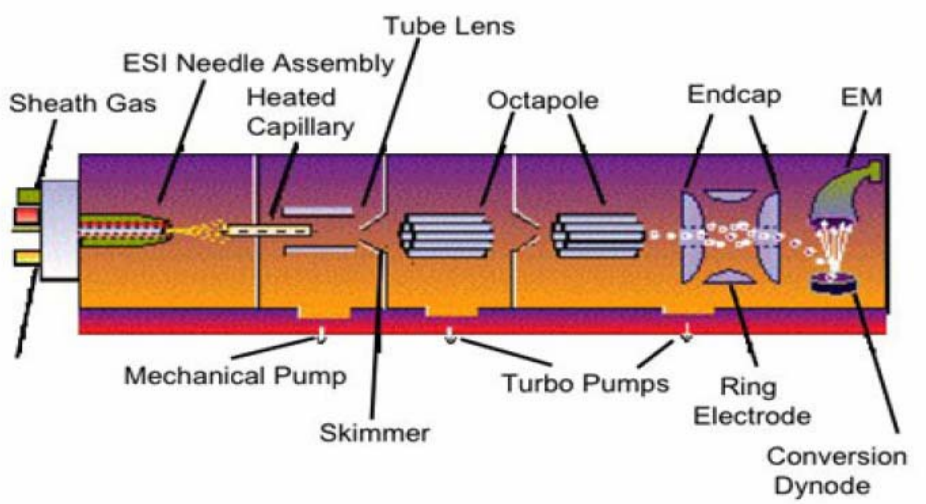


Figure 2.7: Schematic of a quadrupole ion trap mass spectrometer.
The diagram shows the major components to a quadrupole ion trap mass spectrometer. At the start of the diagram the ESI needle assembly (electrospray ionization) provides the ions travel through the mass spectrometer and are detected by the EM (electron multiplier). Figure Provided by Thermo (www.thermo.com).

An ion will be stably trapped depending upon the values for the mass and charge of the ion, the size of the ion trap (r), the oscillating frequency of the fundamental rf (Ω), and the amplitude of the voltage on the ring electrode (V). The dependence of ion motion on these parameters is described by the two dimensionless parameter q_z and a_z, as evident in the formula below [60]. The qz value will determine when m/z ejection takes place.

$$m/z\ eject = 4V/(0.908r^2\Omega^2)$$

For detection of the ions, the potentials are altered to destabilize the ion motions resulting in ejection of the ions through the exit endcap. The ions are usually ejected in order of increasing *m/z* by a gradual change in the potentials. The "stability diagram" depicts the region where radial and axial stability overlap (Figure 2.8). Depending upon the amplitude of the voltage placed on the ring electrode, an ion of a given *m/z* will have a (qz*)* value that will fall within the boundaries of the stability diagram, and the ion will be trapped. If the q z value at that voltage falls outside of the boundaries of the stability diagram (qz = 0.908), the ion will hit the electrodes and be lost. By sequentially increasing the voltage on the ring electrodes, ions trajectories from low *m/z* to high *m/z* are made unstable (Figure 2.8). This "stream" of ions generated from this sequential ejection are focused onto the detector or electron multiplier of the instrument in order to produce the mass spectrum. The initial mass spectrum obtained is what is known as a full scan. The ions observed within the full scan are next selected by their m/z values for isolation and subsequent fragmentation. This selection is accomplished by destabilizing and ejecting all other ions with lower and higher m/z values as described above. The process of selection is essentially gas phase purification of the ion inside the mass spectrometer.

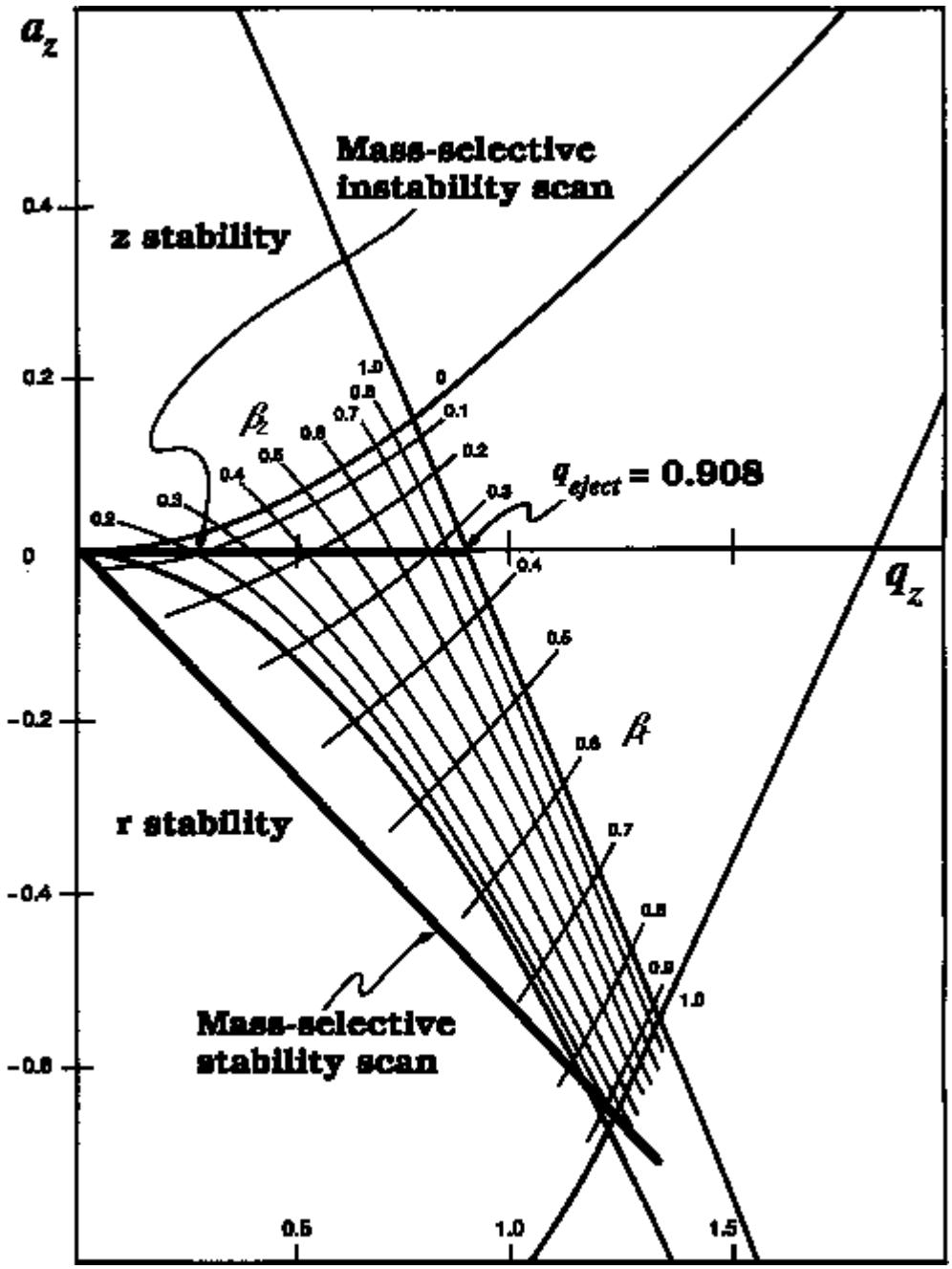


Figure 2.8: Stability diagram for the quadrupole ion trap.
(Figure taken from Karen R. Jonscher and John R. Yates, III, www.ABRF.org)
Diagram showing the regions of stability within the quadrupole ion trap depicted in terms of the operating voltages and frequencies. The important terms on the diagram are the (a) and (q_z) functions. These functions represent stability ranges within the 3D trap.

The selected ion is then excited by increasing its orbital frequency, which causes it to collide with the helium bath gas inside the ion trap mass spectrometer [60]. These repeated collisions with the helium gas cause collision induced fragmentation. After fragmentation occurs, the fragment ions are maintained within the ion trap mass spectrometer. The same process of destabilizing and ejecting ions that occurred in the full scan is performed for the fragment ions thereby producing what is know as an MS/MS or MS^2 spectrum. The entire MS/MS process is repeated for three to four more times for different selected ions within the ion trap before the mass spec returns to a full scan. The process of full scans followed by MS/MS is repeated throughout an entire chromatographic run, creating thousands of MS/MS spectra and their associated parent m/z measurements.

Within this dissertation, 1D-LC-MS/MS was used for all peptide analysis. This methodology is one of the simplest and easiest to implement, and was the reason it was chosen within this body of work. It requires only three major instruments, a low-flow HPLC pump, an autosampler, and the electrospray quadrupole ion trap mass spectrometer (ES-QIT-MS).

1D-LC-ES-MS/MS

All *R. palustris* and *E. coli* protein preparations from each fraction were digested for bottom-up analysis with sequencing grade trypsin added at 1:20 (wt/wt) of protein to enzyme. The digestions were run with gentle shaking at 37 °C for 12 hours. Samples were immediately desalted with an Omics 100 µl solid phase extraction pipette tip (Varian, Palo Alto, CA). All samples were frozen at -80°C until LC-MS/MS analysis.

For all peptide samples, one-dimensional (1D) LC-MS-MS experiments were performed with a Famos/Switchos/Ultimate HPLC System (Dionex, Sunnyvale, CA) coupled to an LCQ-DECA XP Plus quadrupole ion trap mass spectrometer (Thermo Finnigan, San Jose, CA) equipped with a nanospray source as previously described [61]. A 160 minute linear gradient from 100% solvent A (95% H_2O/5% ACN/0.5% formic acid) to 100% solvent B (30% H_2O/ 70% ACN/0.5% formic acid) was employed. For all 1D LC-MS-MS data acquisition, the LCQ was operated in the data dependent mode with dynamic exclusion enabled (repeat count 2), where the four most abundant peaks in every MS scan were subjected to MS-MS analysis. Data dependent LC-MS-MS was performed over a parent m/z range of 400-2000.

Data Analysis

All resulting top-down and bottom-up data sets were analyzed with two methods. In the first method, the well established bottom-up algorithm SEQUEST was used to identify MS-MS spectra with their counterparts predicted from a protein sequence database [62]. The sequence information of the peptide cannot easily be directly interpreted from the MS/MS spectrum due to the complexity of the fragmentation processes. Instead, SEQUEST performs cross correlation comparisons between the observed spectrum and computationally derived spectra from protein and nucleotide databases. The parent mass of the peptide provides a look-up function to find candidate peptide sequences within the potential mass window of the observed parent peptide. The observed MS/MS spectrum is then directly compared to hundreds of potential candidate MS/MS spectra and a best scoring candidate match is made. For all database searches, an *R. palustris* proteome database was used, which contained 4,833 proteins and 36 common

contaminants or the *E. coli* K-12 database plus common contaminants. Distracter databases of other organisms, such as yeast, were used to search against for verification of false positive rates. All resultant output files from SEQUEST were filtered by DTASelect [61] at the 1-peptide, 2-peptides and 3-peptides level with the following parameters: SEQUEST, DeltCN of at least 0.08 and cross correlation scores (Xcorrs) of at least 1.8 (+1), 2.5 (+2) and 3.5 (+3), followed by Contrast [61] for comparison. The DTASelect [61] software can take any number of LC-MS/MS analyses and sort and filter peptide identifications to provide html and text output files of identified proteins, while the Contrast [61] algorithm can compare across multiple outputs from DTASelect [61] for multiple proteomics experiments. The filtering levels used for all searches are considered to be conservative, generally giving less than 1-5% false positive rates at the 2 peptide level depending on the data sample size and the database size.

In the second method, integrated top-down and bottom-up searching was performed with PTMSearch Plus software developed at Oak Ridge National Laboratory (Chapter 6). Output files containing bottom-up data from PTMSearch Plus were filtered by DTASelect [61] at the 2-peptides level with the following parameters: MASPIC [63], scores of at least 23 (+1), 28 (+2) and 43 (+3). These scores were used to give the same approximate 5% false positive rate as with the scores applied for SEQUEST above. The output files containing top-down data were filtered with at least three peaks within the isotopic package, a 3000 Da mass cutoff and a relative abundance of at least 10%. The false positive rate (proteins identified that are not correct identifications) within the top-down searching is considerably higher than the bottom-up methods due to the presence of PTMs. Post translational modifications increase the likelihood of a combination of PTM

masses added to the protein equaling the measured mass being searched. The same scenario is true for false negatives. False negatives are real proteins not identified or not included in the output of identifications due to low scores. The searching algorithms may miss proteins due to a mass with a combination of PTMs giving a better score than the real identification. Another area of concern in top-down data analysis is proteins with good signal to noise ratios and abundant isotopic packages that are not identified. This lack of identification could be due to three reasons. The first reason is degradation and truncation products making the mass significantly different from the predicted masses in the database. Second, a combination of PTMs or unique PTMs leaves the protein unidentified. Finally, missed start calls in the genome annotation process provide wrong protein masses within the database. Due to the false positives, false negatives and no identifications encountered with top-down searching alone an integrated top-down and bottom-up data searching is employed. The integrated searching provides a confident list of proteins from the bottom-up data that the top-down data can be compared against. In the data searching employed within this dissertation a combination of bottom-up peptide data, as well as a top-down intact mass measurement was required for a positive identification.

The PTMSearch Plus program allows for the combined searching of both the top-down and bottom-up data sets; as well as allowing for the searching of a defined set of PTMs (Chapter 6). In the integrated top-down and bottom-up data searches a standard set of PTMs were searched for including: methylation, acetylation, N-terminal methionine truncation, and disulfide bonds (restricted to top-down data). Less common PTMs such as uridylylation were searched individually. All data outputs generated are

manually inspected and then compared using Microsoft Access (Microsoft Corp., Redmond, WA).

Chapter 3

Extension of FTICR-MS Methodology for Proteins and Peptides: Advanced Charge State Determination and Alternative Fragmentation Approaches

Data presented below is in preparation for submission or in press

Heather M. Connelly and Robert L. Hettich. Comparison of MSAD and SORI-CAD for Peptides and Peptide Mixtures Using Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Journal of the American Society of Mass Spectrometry, In final preparation* (2006). *All MS sample preparation, experiments and data analysis were performed by Heather M. Connelly.*

David L. Tabb, Manesh B. Shah, Michael Brad Strader, Heather M. Connelly, Robert L. Hettich, Gregory B. Hurst. Determination of Peptide and Protein Ion Charge State by Fourier Transformation of Isotope-Resolved Mass Spectra. *Journal of the American Society of Mass Spectrometry, Accepted, In Press* (2006). *All FTICR-MS sample preparation and experiments were performed by Heather M. Connelly.*

Introduction

In the process of this dissertation work, a need for fundamental advancements in the analysis of proteins and peptides was essential. Two areas of particular interest were better methods for determination of charge states for large proteins and advanced protein fragmentation methods with FTICR-MS. Both of these areas were examined to improve the overall experimental platform for identifying intact proteins and their associated PTMs.

The first area of development was the robust determination of charge states for large proteins. Generally, the determination of the charge state for an ion with FTICR-MS is straight forward if the spectrum is sufficiently resolved to distinguish peaks in the isotope packet for the ion. FTICR mass spectrometry provides very high resolution and accuracy because of the accuracy with which it is possible to measure the frequency of

ion cyclotron motion in the Penning trap [64]. This resolving power enables acquisition of mass spectra of electrosprayed intact protein ions with resolved isotopologues. Due to the unparalleled mass resolution and accuracy of the FTICR-MS, this is usually not a problem for proteins under electrospray conditions in which the analyte concentration and ion detection parameters can be optimized. A more challenging scenario is presented for situations in which these carefully controlled conditions are not possible, such as LC-FTICR-MS measurements. In this case, the signal quality is compromised and the direct measurement and resolution of charge states from intact proteins is much more difficult.

Due to the difficulties encountered with ion trap measurements and charge state determination, Dr. David Tabb (ORNL post doctoral research associate with the OBMS group) developed an automated method for determining charge states from high-resolution zoom scans within the linear ion trap. Further, we decided this method could be applied to LC-FTICR-MS measurements and charge state determination, in order to perform an automated method for determining charge states from high-resolution mass spectra. Fourier transforms of isotope packets from high-resolution mass spectra are compared to Fourier transforms of modeled isotopic peak packets for a range of charge states. The charge state for the experimental ion packet is determined by the model isotope packet that yields the best match in the comparison of the Fourier transforms.

The second area targeted for FTICR-MS development was the evaluation of proteins and peptide fragmentation methods within the FTICR-MS. A number of tandem mass spectrometry methods are employed to dissociate peptides and proteins using a FTICR-MS instrument; these include collision activated dissociation (CAD), surface induced dissociation (SID), electron capture dissociation (ECD), and multiphoton

infrared photodissociation (IRMPD). Traditionally MS/MS on a FTICR-MS has been accomplished with sustained off-resonance irradiation collisionally activated dissociation (SORI-CAD). This method makes use of the off-resonance excitation of the parent ions being investigated [65]. SORI-CAD is capable of producing high fragmentation efficiency with relatively simple implementation [65]. However, the need for manual individual parent ion selection, the low duty cycles required, and the delay for pump down after the introduction of pulsed collision gas make SORI-CAD difficult to use on complex mixtures. In contrast, electron capture dissociation uses low-energy electrons to neutralize the charges on the protein, producing cleavage of the amide bond to form c and z ions [66]. Surface induced dissociation (SID) allows for large amounts of energy to be deposited into a molecule in a very short amount of time [67]. Also, SID does not have the problematic introduction of collision gas, as with SORI-CAD, and has been implemented successfully to a FT-ICR for the study of biological molecules by Laskin *et al*. [67]. This method, although successful, can create the problem of charge neutralization and requires specialized equipment and implementation that may not be readily available in most instances. IRMPD offers a method of fragmentation where no single frequency excitation is required and the ions of all m/z values are dissociated at the same time [68]. This method has been demonstrated useful with biomolecules [68]. However, IRMPD is not always universally available, making another method capable of dissociation of all m/z values at the same time desirable.

Recently, new techniques for dissociation have been employed using the rf-only multipole within the external source of an ESI FT-ICR-MS including: multipole storage assisted dissociation (MSAD) [69, 70], “ion thrashing”[71], and photon-induced

dissociation known as external IRMPD [72]. Previously, it was demonstrated that electrospray generated ions can be externally accumulated in an rf-only multipole prior to mass analysis by a FT-ICR [73]. Multipole storage assisted dissociation (MSAD) was first observed when ions were accumulated in the rf-only multipole for an extended period of time [74]. During a MSAD experiment there is not any apparent contact with the rods of the hexapole that could generate a surface induced dissociation. Space charging in the hexapole seems to push the ions out radially allowing them to obtain rf from the rods generating ions with higher kinetic energy [72,75, 76, 77]. The excited ions generated are then able to be fragmented with the background gas molecules in the hexapole (air at ~10^{-5} torr), making it a form of CAD [78]. MSAD allows for ion activation and dissociation simultaneously with ion accumulation and no collision gas is introduced into the analyzer cell, so no pump down period is needed creating a more efficient method [78]. Like IRMPD, the MSAD method provides an effective way to accomplish dissociation on all m/z values at once, but does not provide a way to perform targeted fragmentation. Also, this method is quite accessible unlike IRMPD since most FT-ICR instruments are equipped with a linear ion trap at the interface of the electrospray ionization source to the FTICR cell.

The level of fragmentation observed using MSAD is a function of hexapole accumulation time, dc off set voltage applied, and concentration of sample being used [75]. The dc offset voltage controls the depth of the electrostatic axial well [77,79]. The larger the dc offset voltage that is applied, the greater the capacity of the ion reservoir within the hexapole which allows for more space charging and dissociation by MSAD [79]. Extended ion accumulation times provide a larger population of ions with in the

multipole facilitating in the space charging and ion oscillation at higher amplitudes [80]. Also important is the total sample concentration; varying sample concentration can require the need for different accumulation times to induce dissociation in MSAD experiments.

This method has been used to successfully generate fragment ions in intact proteins, although, little has been reported on the efficacy of using MSAD on peptides. A study by Haselmann et al. compared the effects of SORI-CAD, MSAD, and ECD on a single peptide and found fragment peaks more abundant with MSAD, when compared to SORI-CAD [81]. However, the Haselmann et al. study gave good preliminary results, an exhaustive MSAD and SORI-CAD comparison of peptides and peptide mixtures was not performed until this study. In this work, we report on the efficacy of using MSAD instead of SORI-CAD on single peptide solutions, simple peptide mixtures and peptide solutions from tryptic digest of intact proteins to provide in-depth data on fragmentation patterns, ion series generated, and spectral complexity.

Methods and Materials

LC-FTICR-MS of Intact Proteins for Charge State Determination

Five proteins (ubiquitin, chicken lysozyme C, bovine ribonuclease A, bovine carbonic anhydrase II, and bovine beta lactoglobulin-B) were dissolved in HPLC grade water to give a final concentration of 1 mg/mL of each protein, and diluted as required for the analysis. All capillary HPLCFTICR experiments were performed with an Ultimate HPLC (LC Packings) coupled to an IonSpec 9.4 T FTICR-MS (Lake Forest, CA) mass spectrometer equipped with an Analytica electrospray source. A Vydac 214MS5.325 (Grace-Vydac, Hesperia, CA) C4 reverse phase column (300 m i.d. X 250 mm, 300 Å

with 5 m particles) was directly connected to the Analytica electrospray source with 100 micron i.d. fused silica tubing. Injections of 30 μg of total protein were made onto a 100 μl loop. The flow rate was 4 μL/min, with a 75 min gradient going from high water (95% water, 5% acetonitrile, 0.5% formic acid) to high organic (95% acetonitrile, 5% water, 0.5% formic acid). All mass spectra were acquired with a 2 s hexapole ion accumulation time; 2 scans were signal averaged, 1024 K data points were acquired, and 2 zero fills were performed. The Hann window was used for apodization. Mass resolving powers of 35,000 to 120,000 FWHM were achieved. Mass calibration was performed externally using an ubiquitin protein standard, providing approximately 10–50 millidalton accuracy. Mass spectra were viewed via the Omega 8 instrument control software provided by IonSpec. The most abundant isotopic mass (MAIM) for each protein was computed [82], and a spreadsheet calculated the m/z ratio corresponding to each charge state. To compute the MAIM values the most abundant isotopic mass within the isotopic package was compared to a calculated most abundant mass within the isotopic package. To obtain the calculated MAIM, the sequence of the protein was input in to the PAWS [84] software to obtain the number of each molecular atom present for the sequence. Once the molecular atoms were obtained, they were input into the Exact Mass Calculator, provided as part of the IonSpec software package, in order to determine the calculated most abundant isotopic mass within the isotopic package. Three mass spectra from the LC-FTICR-MS data, containing charge state packets for the five proteins, were chosen for charge state analysis. The FTDocViewer “Isotope Clusters” feature displayed the isotopic packets from each spectrum along with the assigned charge state(s). A beta-version of IonSpec’s PeakHunter algorithm (version 0.0.24) was then used to assess charges for the

same spectra. Scripts were developed for examination of the data in external software. The mass spectra from the IonSpec instrument are extracted to an MS1 file [83] by "MakeMS1", a Visual Basic Script for FTDocViewer in the Omega8 instrument control software. The isotopic packets for proteins ranging in charge from z = 5 to 30 are modeled, and FFTs of these charge models are stored. The observed mass spectra are read into memory by the "Tact" algorithm, C ++ software created at ORNL for analysis of FTICR data from intact proteins. The software identifies the set of nonoverlapping one m/z-wide windows containing the highest intensity within each mass spectrum. The FFT of each one m/z-wide window is computed, and the charge model FFT that best matches the FFT of the observed spectrum (in terms of normalized dot product score) is stored as the charge state for that packet.

Methods for MSAD Fragmentation

Eight synthetic peptides along with angiotensin I, angiotensin II, Neurotensin, Bradykinin, Des-Arg Bradykinin, Thr-Bradykinin, and Meth-Enkephalin were used as purchased from Sigma-Aldrich (St. Louis, MO) without further purification. Each of the peptides and protein solutions were prepared in 50/50 Acetonitrile/water: 0.1 % Acetic Acid to a total concentration of 10 μM**.** Acetonitrile and HPLC grade water were purchased from Burdick and Jackson (Muskegon, MI). Acetic acid (99.9 %) was from Sigma-Aldrich (St. Louis, MO).

Mixtures of Angiotensin I, Meth-Enkephalin, synthetic peptides 3, 4, 6, and 7 were made using 10μM concentration solutions at a ratio of 1:1 and also peptide was mixed at a 1:100 ratio to the other 5 peptides. Bovine Serum Albumin (Sigma-Aldrich, St. Louis, MO) and Horse Apomyoglobin (Sigma-Aldrich, St. Louis, MO) were

denatured with 6M Guanidine and 5 mM DTT at 60^{O}C for 1 hour and then diluted in 50 mM Tris (pH 7.5)/ 5 mM $CaCl_2$ to obtain a final Guanidine concentration of 1 M. Sequencing grade trypsin (Promega, Madison WI) was added at a concentration of 1:50 and allowed to digest for 16 hours. Trypsin was then added a second time at a concentration of 1:50 and digested for another 6 hours, followed by a final reduction step with 10mM DTT for 1 hour. Samples were immediately desalted with a C18 Sep-Pak (Waters, Milford MA) and concentrated by centrifugal evaporator (Savant Instruments, Holbrook, NY). Samples were diluted in 50:50:0.1 ACN:H_2O:HOAc to a total concentration of 10 μM.

ES-FT-ICR mass spectra were all acquired with an IonSpec (Lake Forest, CA) 9.4-Tesla (Cryomagnetics Inc., Oak Ridge, TN) HiRes electrospray Fourier transform ion cyclotron resonance mass spectrometer. A Harvard syringe pump set at a flow rate of 1.75 μL/min was coupled to an Analytica electrospray source (Analytica of Branford, CT). After generation, ions were accumulated in an external hexapole and transferred into the high-vacuum region with a quadrupole lens system. Detection then followed in the cylindrical analyzer cell of the mass spectrometer. Calibration of the mass spectrometer was accomplished externally with ubiquitin resulting in a mass accuracy of ±3 ppm and resolutions of 50,000-160,000 (FWHM) for peptides.

To perform ion collisional dissociation an ion of interest was isolated from a peptide within the analyzer cell of the mass spectrometer and then accelerated into a nitrogen target gas under sustained off-resonance irradiation collision-activated dissociation (SORI-CAD). An rf pulse set at ~ 1KHz lower in frequency applied for 2 seconds at an amplitude range of 2-5 volts was used for the ion excitation in SORI-CAD

experiments. During the ion excitation step a pulsed valve was used to admit the nitrogen collision gas into the high vacuum region to a maximum pressure of about 5 X 10^{-6} Torr. Prior to ion detection the base pressure was returned to 6X10^{-10} Torr.

Under normal mass spectrum conditions the dc voltage in the rf-only hexapole, located at the interface of the electrospray source and the FT-ICR cell, is at -3.5v with an ion accumulation time of one to two seconds. These experimental parameters generate multiply charged molecular ions with virtually no fragmentation. When performing a MSAD experiment these conditions are altered to facilitate dissociation within the hexapole. The dc offset voltage is decreased to -7 to -11 volts and accumulation time is increased to 4-5 seconds creating extensive fragmentation of the ions within the rf-only hexapole. After dissociation and injection into the ICR cell, ion detection followed. Since no parent ion isolation, activation, or pump down delays from collision gas addition were needed, overall scan functions for MSAD experiments generally took 2 to 6 s per transient acquired. Each spectrum obtained was comprised of 2 co-added transients acquired at 1024K data points. Deconvolution of product ion spectra to a zero charge state was accomplished with the IonSpec deconvolution software.

All peptide samples were tested at a number of different offset voltages and accumulation times. It was found that a 5 second accumulation and -11 V offset voltage were the optimal conditions to produce MSAD fragmentation in all samples, therefore, all MSAD data presented have these conditions. Also important is the total sample concentration; this is why all peptide samples were kept at 10uM for both the SORI-CAD and MSAD experiments thus preventing the need for different accumulation times to induce dissociation in MSAD experiments [70]. Due to the extensive fragmentation

observed in a MSAD experiment, a high resolution instrument, such as an FT-ICR, is needed to resolve these complex spectra.

Data analysis was accomplished with the ProteinInfo function of the PROWL website provided by the Rockefeller institute [84]. PROWL is a protein analysis website that enables users to perform mass calculations, mass spectrometry fragmentation, and insilico digest of proteins. These functions within the PROWL website allow for the comparison of experimental data to calculated data for the protein. Both CAD and MSAD fragments, generated in the mass spectra, were analyzed under the mass spectrometry fragmentation function to assign fragmentation patterns. Manual inspection was used to verify all PROWL results and to search for additional identification of internal fragments from the loss of water, which is due to PROWL only assigning ammonia loss. For the simple peptide mixture, the MSAD spectrum of the mixture was compared to the individual MSAD spectrum for each component in order to identify which fragment ions were generated from each individual peptide component present in the mixture. These matching MSAD fragments could then be assigned identifications using PROWL. Fragment ions identified from BSA and Apomyoglobin with PROWL were only reported to two decimal places by the program. To obtain a more accurate mass to compare to the FT-ICR MSAD fragments for tryptic digest of both BSA and Apomyoglobin, the PAWS [84] program was used to determine the atom composition, of MSAD fragment ions. Following the determination of atom composition the IonSpec exact mass calculator was used to calculate the exact mass of the fragment ion that could be compared to the mass of the MSAD fragment ion observed in the spectrum.

Results for Charge State Determination of Intact Proteins

Charge Measurement for Intact Protein Isotope Clusters in FTICR Spectra

The resolution of FTICR mass spectrometry makes it possible to apply the developed charge state determination technique to small sections of normally-acquired mass spectra. The Tact software, developed at ORNL by Dr. David Tabb, for intact protein identification from FTICR data was adapted to find isotopic packets in collections of mass spectra and perform charge state assignments by FT. The FFTs of these packets were compared to FFTs of modeled isotopic packets in order to determine charges by moving across the experimental isotopic package while comparing how well the overlapping experimental isotopic peaks match. Because most proteins adopt multiple charge states under electrospray conditions, multiple isotopic packets of known charge are available to test charge state detection algorithms. Three mass spectra from a liquid chromatographic separation interfaced via electrospray with the FTICR were examined; mass spectrum 10 included charge packets for ribonuclease A, mass spectrum 23 showed the presence of ubiquitin and lysozyme, and mass spectrum 42 gave evidence for beta lactoglobulin and carbonic anhydrase. Table 3.1 compares the performance of IonSpec's "FTDocViewer" and "PeakHunter" software to that of Tact for charge state inference. Each charge determination reported from Tact is the top-scoring match.

Table 3.1: Automated protein charge state assignments from FTICR data

			*Charge State Assignment**			
Charge	*MAIM m/z*	*Intensity*	*FTDoc Z*	*PeakHunter Z*	*Tact Z*	*Tact Score*
Ribonuclease A, Mass Spectrum 10						
7	1956.04	0.05	No Call	No Call	5	0.46
8	1711.66	11.35	10, **8**	10,4,4	9	0.58
9	1521.59	21.14	10, **9**	1, 2, **9**	11	0.49
10	1369.53	7.15	**10**	**10**	**10**	0.75
11	1245.12	5.80	**11**, 1	**11**	**11**	0.87
12	1141.44	0.83	3, **12**	**12**	**12**	0.91
Ubiquitin, Mass Spectrum 23						
5	1713.92	0.06	2, 4	No Call	**5**	0.76
6	1428.44	4.80	**6**	**6**	**6**	0.86
7	1224.52	26.74	**7**	**7**	**7**	0.83
8	1071.58	19.30	**8**	**8**	**8**	0.84
9	952.62	15.98	**9**	**9**	**9**	0.97
10	857.46	5.53	**10**	**10**	**10**	0.97
11	779.60	2.67	**11**	**11**	**11**	0.92
12	714.72	0.49	**12**	**12**	**12**	0.89
Lysozyme, Mass Spectrum 23						
9	1590.87	1.16	**9**, 2	**9**, 4	**9**	0.76
Beta Lactoglobulin, Mass Spectrum 42						
11	1662.67	0.09	8	No Call	18	0.47
12	1524.20	0.10	8, 6, 2	**12**, 3, 3	6	0.37
13	1407.03	0.40	2, **13**	14	**13**	0.49
14	1306.60	3.61	2, **14**	1, **14**	**14**	0.50
15	1219.56	2.31	1, 14, **15**	**15**, 2	**15**	0.59
16	1143.40	0.09	2	16	**16**	0.65
17	1076.20	0.14	No Call	29, 7	27	0.34
Carbonic Anhydrase, Mass Spectrum 42						
21	1383.08	0.08	4	3	30	0.31
22	1320.26	0.11	5	4	**22**	0.31
23	1262.90	0.38	No Call	8, 8	**23**	0.26
24	1210.32	0.30	1	12, 12	8	0.29
25	1161.95	0.17	1	10, 5	**25**	0.48
26	1117.30	0.16	2, 4	**26**	**26**	0.38
27	1075.95	0.14	No Call	29, 7	**27**	0.34
28	1037.56	0.07	No Call	No Call	**28**	0.54
29	1001.82	0.10	No Call	**29**	**29**	0.63
30	968.46	0.07	No Call	No Call	**30**	0.40

*Correct charge assignments are shown in **bold font**.

While Table 3.1 also lists the Tact score for each top-ranking assignment, it is important to emphasize that these scores are not used in an absolute sense, but rather for ranking matches for each isotope packet. That is, there is no absolute threshold score above which a charge state assignment is accepted. Instead, the reported charge state assignment is simply that with the highest Tact score. While Tact reported only one charge assignment for each peak packet, FTDocViewer and PeakHunter can report multiple charge assignments for each set of isotopic peaks, giving them a better chance of randomly hitting the correct charge, but reducing their specificity.

The isotopic packets in scan 10 for ribonuclease A were intense, but also contained additional isotopic packets near the most intense packets, suggesting that other forms of the protein were also present. Perhaps, because of these additional packets, the two most intense packets, corresponding to the $z = 8$ and $z = 9$ charge states of the protein, resulted in multiple charge state calls by FTDocViewer and PeakHunter and an incorrect charge assignment by Tact. For less intense isotope packets corresponding to higher charge states, Tact and PeakHunter yielded correct results, while FTDoc returned multiple possible charges for the $z = 11$ and 12 states. Scan 23 included isotopic packets for ubiquitin and lysozyme. Ubiquitin's packets for $z = 7$ through 9 were the most intense, and they were more than an order of magnitude more intense than lysozyme's sole isotopic packet at $z = 9$. All of these packets were assigned the correct charge by all three algorithms. The $z = 5$ charge state for ubiquitin, however, was called correctly by only the Tact algorithm. This isotopic packet was the least intense to be assigned a correct charge in this collection of mass spectra. Scan 42 comprised a much greater challenge. β- lactoglobulin and carbonic anhydrase both contributed isotopic packets, but

β-lactoglobulin's $z = 14$ charge state was approximately 10-fold more intense than the most intense carbonic anhydrase isotopic packet. FTDocViewer and PeakHunter both yielded multiple charge state calls for many isotopic packets. In several cases, the packets for carbonic anhydrase were not assigned charges by these algorithms, presumably because these low-intensity peaks were not easily centroided. Tact, however, was able to assign four consecutive correct charge states for β-lactoglobulin. For carbonic anhydrase, only Tact was able to achieve any consistency, assigning eight of the 10 charge states correctly. Scan 42 demonstrates that FFT is particularly powerful for inferring charge states from noisy signals of low intensity. Overall, Tact performed comparably to FTDoc-Viewer and PeakHunter in determining charge states from ion packets of moderate intensity and signal-to noise ratio, while providing an improvement for low abundance, noisy isotope packets. It is important to keep in mind that, in the context of this LC-MS experiment, the proteins were available during only a limited time for MS data acquisition during elution of a peak in a liquid chromatography separation. As such, the optimal performance factors for high-resolution mass measurement must be compromised somewhat to accommodate the shorter time frame for ion detection (i.e., few scans and fewer data points for the transient signal). The exquisite resolution possible for FTICR instruments, along with FTDocViewer and PeakHunter processing algorithms, enables accurate determination of charge states for proteins up to at least 60 kDa under direct infusion conditions, where protein concentrations and ion accumulation and detection parameters can be optimized. Accurate charge state determination is a critical component for computational programs such as THRASH [85], which seek to combine

this information with isotopic abundance in order to permit comparisons between measured and predicted mass spectra for molecular identifications.

Results for the Comparison of MSAD and SORI-CAD for Peptides and Peptide Mixtures

SORI-CAD and MSAD for Single Peptides

A range of peptides were examined with MSAD in this study in order to determine the general utility of the technique, and to make a general comparison with SORI-CAD. It has been previously demonstrated that proteins have similar fragmentation patterns in both low energy MSAD experiments and in SORI-CAD [69, 77]. However, in this study, high energy MSAD (based on experimental observation) was used to take advantage of the unique property of MSAD, which is the ability to put high amounts of collisional energy into the peptide. This high energy MSAD was performed by simply elongating the accumulation time to 4-5 seconds from 2 seconds and adjusting the magnitude of the dc offset to -11V from -7V for all peptides. These high energy MSAD experimental parameters were compared with standard SORI-CAD conditions. SORI-CAD experiments were accomplished with a rf pulse, set at ~ 1KHz lower in frequency than normal mass spectra acusition, applied for 2 seconds at an amplitude range of 2.8-4.0 volts to ensure complete dissociation of the peptide.

In this study, we have conducted MSAD and SORI-CAD experiments with peptides that exhibit a wide range of diversity in their amino acid sequence, molecular weight and post translational modifications (Table 3.2). The MSAD spectrum and SORI-CAD spectrum are shown for both synthetic peptide 1 and Bradykinin (Figure 3.1).

Table 3.2 Name, sequence and molecular weight of all peptides used.

	Sequence	MW
synthetic 1	Acetyl-RAYIFAVR-OH	1036.6
synthetic 2	AQTERKSGKRQTER	1673.9
synthetic 3	GKAKVTGRWK	1129.7
synthetic 4	VHLTPVEK	922.1
synthetic 6	MEMKKVLNS	1079.3
synthetic 7	FLEEI	649.7
synthetic 8	YIGSR	594.7
Angiotensin 1	NRVYIHFPHL	1296.5
Angiotensin 2	NRVYIHPF	1046.2
Bradykinin	RPPGFSPFR	1060.2
Thr-Bradykinin	RPPGFTPFR	1074.2
Des-Arg Bradykinin	PPGFSPFR	904
Meth-Enkephalin	YGGFM	573.7
Neurotensin	ELYEDKPRRPYI	1672.9

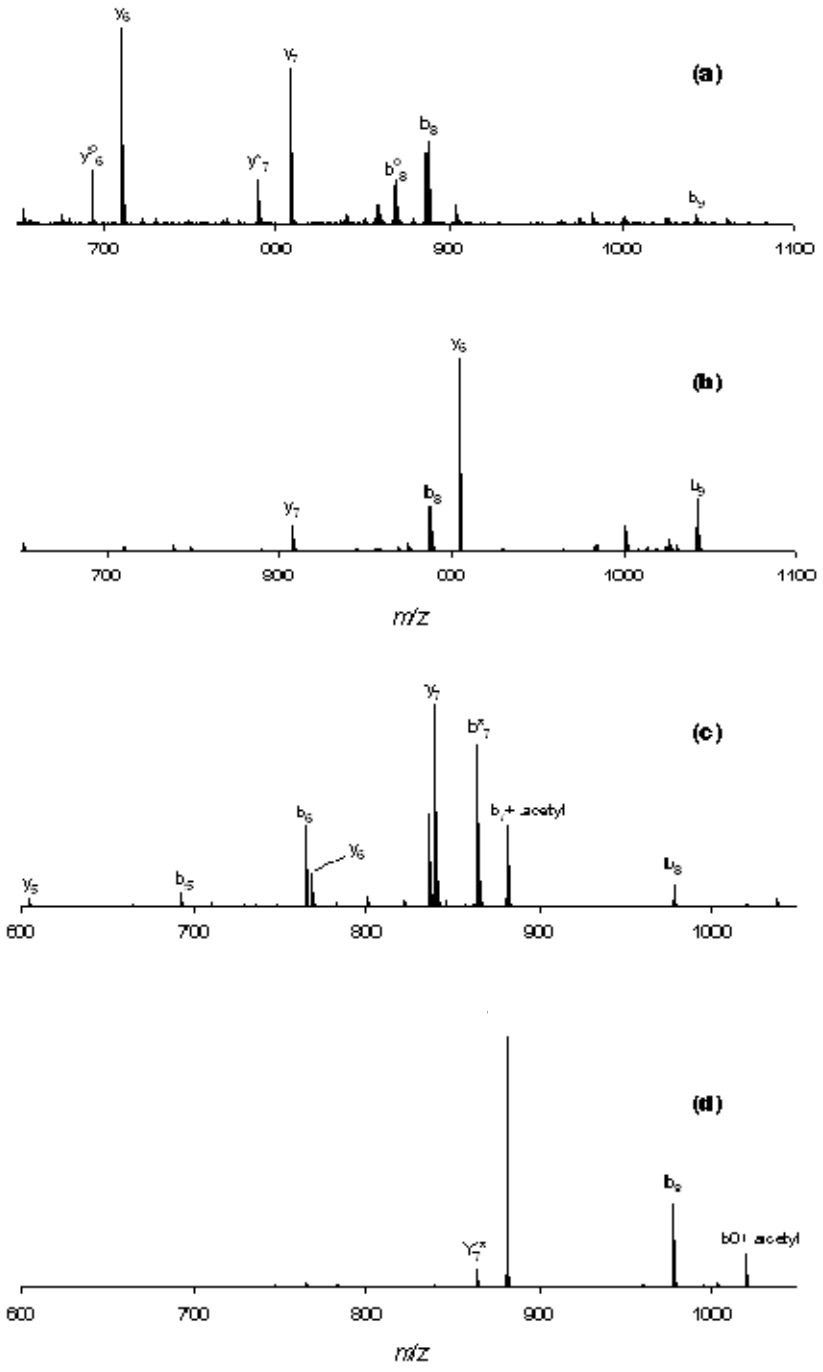


Figure 3.1: B and Y ion labeled MSAD and SORI-CAD spectrum for Bradykinin and Synthetic peptide 1. The loss of water labeled with (*) and the loss of ammonia labeled with (o). (A) MSAD spectrum of Bradykinin (B) SORI-CAD spectrum of Bradykinin (C) MSAD spectrum of Synthetic 1 (D) SORI-CAD spectrum of Synthetic 1.

In the Bradykinin MSAD spectrum (Figure 3.1A) the loss of water and ammonia occurs with the prominent b or y ion. In the case of the y_7 ion within the MSAD spectrum (Figure 3.1A) the loss of water is most likely from the serine side chain contained in the sequence of the fragment ion. In contrast to the MSAD spectrum, the SORI-CAD spectrum (Figure 3.1B) for Bradykinin has the same prominent b_8, b_9, y_7, and y_6 ions without the loss of water and ammonia. To simulate fragmentation of peptides with post translational modifications, synthetic peptide 1 was chosen because it has an acetyl group on the N-terminal. In the MSAD spectrum of synthetic peptide 1, the acetyl group is maintained only on the b_7 ion and not on any other b ions or the y-ion series (Figure 3.1C). In comparison, the SORI-CAD spectrum has the acetyl group maintained on the fragment ion b_8 (Figure 3.1D). Both the MSAD fragment ions for synthetic peptide 1 contain the acetyl group on one ion; the SORI-CAD spectrum also has the acetyl group on the prominent b ion within the spectrum.

Both SORI-CAD and MSAD experiments provided identifiable peptide fragmentation patterns when searched using the PROWL website [84]. Comparisons of identifiable fragments from both methods reveal a more extensive fragmentation pattern in the MSAD spectra, with 10 out of 15 peptides showing more identifiable fragment ions than seen with SORI-CAD (Figure 3.2).

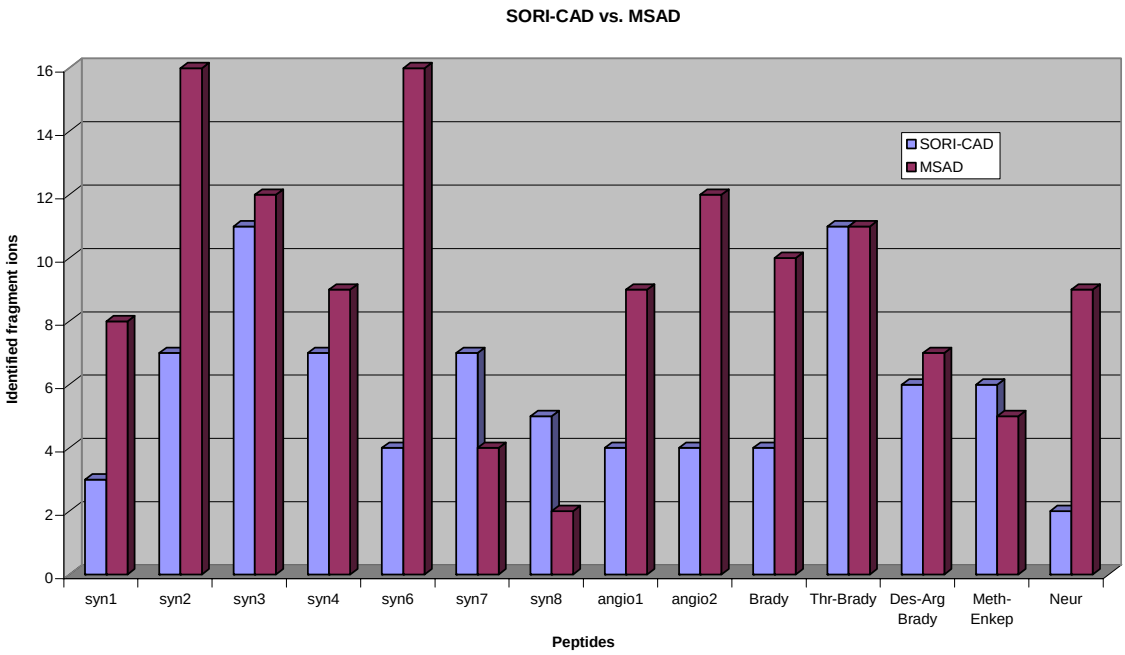


Figure 3.2: Comparison of MSAD and SORI-CAD fragment ion identifications for all 14 peptides.

MSAD experiments produce more fragment ions than SORI-CAD, although, the fragment ions that are produced in SORI-CAD are more easily identified and give a spectrum where almost all fragment ions are b and y ions, without the internal fragments and loss of water and ammonia. Table 3.3 shows the most abundant MSAD and SORI-CAD fragment ions for each peptide. On average, there are 2 to 4 fragment ions observed in the SORI-CAD spectrum; in comparison, there are 4 to 10 fragment ions in the MSAD spectra (Table 3.3). When comparing the MSAD and SORI-CAD fragment ions, the predominant fragment ions in the spectra of both methods are often the same. This can be seen in the comparison of MSAD and SORI-CAD spectra for synthetic peptide 4 where both methods have a prominent b_8 and b_7 ions (Table 3.3). The differences in the amount of abundant fragment ions in the case of synthetic peptide 4 is the b_8 and b_7 fragment ions in the MSAD spectrum are also accompanied by fragment ions with a loss of water, ammonia, or both. Another difference in the number of abundant fragment ions observed between MSAD and SORI-CAD spectra are the amount of internal fragments. This is demonstrated in synthetic peptide 6, where some of the same predominant b and y ions within the MSAD spectrum have five additional internal fragment ions (Table 3.3). MSAD experiments allow for the dissociation of all parent ions within the rf-only hexapole without pre-selection or isolation of parent ions, as required with SORI-CAD experiments. This simultaneous dissociation of parent ions gives a (b and y) series of ions although, this series has numerous internal fragment ions with water and ammonia loss. The internal fragment ions observed are possibly occurring from the layering sequential fragmentation, where a y-ion is formed followed by another fragmentation event that fragments the y-ion, giving a b-ion creating an internal fragment that has both a

Table 3.3: Most abundant fragment ions from MSAD and SORI-CAD

	SORI-CAD		MSAD	
Peptide	mass	ID	mass	ID
synthetic 1	1019.5735	b8+acetyl	976.5490	b8
	976.5340	b8	881.4810	b7+acetyl
			863.4708	b*7
			839.4692	y7
			767.4292	y6
			763.3966	b6
			692.3616	b5
			604.3691	y5
synthetic 2	1656.9046	b14	787.4061	y13b8, y12b9
	1499.8229	b13	700.3772	y7b13
	1352.7613	b12	642.3486	y*12b8, y12b*8
	1245.0000	y10	625.3229	y*12b*8
	300.6062	b3		
synthetic 3	874.5290	y7	1130.6771	y10
	745.4103	y6	797.4930	b8
	646.3574	y5	797.4831	b8
			745.4221	y6
			728.4116	y*6,
			613.3707	y8b8, y*5b10
synthetic 4	904.5306	b8	922.5342	y8
	775.4280	b7	904.5360	b8
	572.3182	y5	886.5207	b^o8
	471.2693	y4	775.4272	b7
			757.4186	b^o7
			740.3910	b*^{o}7
			685.4032	y6
			667.3974	y*6
			649.3903	y*^{o}6
synthetic 6	859.4900	b7	1060.5515	b9
	819.4605	Y7	973.5278	b8
	745.4040	b6	859.4911	b7
	688.4163	y6	714.4148	y7b8
			583.3723	y7b*7, y*7b7
			542.3089	y5b9, y*5
			697.4080	y^o7b8
			679.3906	y^o7b^{o}8
			649.3341	y5
synthetic 7	631.3324	b5	631.3158	b5
	518.2440	b4	649.3341	y5
synthetic 8	576.3034	b5	595.3240	y5
Angiotensin 1	1295.7120	y10	1277.6543	b10
	1027.5310	b8	1181.6592	y9
	930.5140	b7	1027.5335	b8
	512.2739	y4	895.4815	y8b9
			783.4191	b6
			765.4006	b^o6

Table 3.3: Continued

Peptide	SORI-CAD		MSAD	
	mass	ID	mass	ID
Angiotensin 2	1027.5313	b8	1045.5322	y8
	931.5180	y7	1027.5125	b8
			784.4000	b6
			765.3886	b *6
			669.3738	y7b6
Bradykinin	1042.5724	b9	886.4480	b8
	904.4668	y8	869.4369	b^o8
	886.4446	b8	806.4099	y7
	806.4059	y7	789.3976	y*7
			709.3550	y6
			992.3518	y^o6
Thr-Bradykinin	918.4921	y8	1074.5695	y9
	554.2995	b5	899.4847	b8
			820.4328	y7
			802.4019	y^o7
			803.4057	y*7
			723.3864	y6
			705.3649	Y^o6
Des-Arg Bradykinin	886.4558	b8	806.3939	y7
	729.3487	b7	789.3827	y*7
	709.3772	y6	709.3678	y6
			652.3238	y5
			634.3130	y^o5
Meth-Enkephalin	555.2069	b5	585.1631	y6b8, y*7b7, y7b*7
	424.1742	b4	573.2314	y5
Neurotensin	660.3990	y5	633.8616	y11b7
			556.9702	y*6b11, y6b*11

y and b end. This effect could be due to the ions multiple pass in the z-direction in the hexapole as proposed by Pan et al. [77]. It has also been demonstrated that the proportion of fragment ions increase with pressure [86]. The effect of pressure and space charging in the rf-only hexapole will lead to a smaller mean free path for the ion and higher oscillation amplitudes creating more kinetic energy and therefore leading to the creation of more internal fragment ions [86]. Unlike MSAD, SORI-CAD takes place in the ICR cell where there is an order of magnitude difference in pressure from the hexapole creating ions with a larger mean free path that experience less space charging thereby leading to lower oscillation amplitudes and no observed internal fragment ions. Due to the nature of MSAD, where it has been suggested that the lowest energy process is selected for dissociation [75], the loss of ammonia and water is often seen as compared to SORI-CAD where there is rarely a loss of water or ammonia. The loss of water in a MSAD experiment is mainly coming from the side chain of the amino acids, such as threonine and serine (since it is a low energy requirement dissociation) [87], while the loss of ammonia is primarily coming from the the N-terminal. SORI-CAD also gives b and y ions, but there are rarely water and ammonia loss and also no internal fragment ions present.

MSAD for Simple Mixtures

Even though MSAD can not isolate a parent ion for fragmentation, a combination of peptides, from the single peptide standards, was used to provide individual peptide identifications, as well as examine sensitivity and dynamic range for the mixtures. Six individual peptide standards (synthetic 3,4,6,7, angiotensin1, meth-enkephalin) were chosen and mixed at a 1:1:1:1:1:1 concentration for all six peptides in order to form a

simple peptide mixture. A mass spectrum was obtained for the mixture followed by a MSAD spectrum (Figure 3.3). These two spectra were compared to see if any parent ions remained and to examine the extent of fragmentation for the individual peptides within the mixture. Each parent peptide within the mass spectrum (Figure 3.3A) was assigned a color label which was used to label corresponding fragment ions within the MSAD spectra (Figure 3.3B). As indicated by the dashed lines, some of the parent peptides did remain in the MSAD spectrum for angiotensin1, synthetic 3 and synthetic 4 (Figure 3.3).

The major fragment ions in the mixture MSAD spectrum were generally the same fragment ions found in the MSAD spectrum for the individual peptide. This can especially be seen for angiotensin 1 where all of the abundant fragment ions produced in the mixture MSAD spectrum are the same as those found in the MSAD spectrum of angiotensin 1 (Table 3.3, Figure 3.3B). However, Meth-Enkephalin did not produce any major fragment ions within the MSAD spectrum, as seen when fragmented as a single peptide, but this could be due to the small size of this peptide (MW=573.7). The similar MSAD fragmentation patterns for individual peptides, as well as in the mixture, make it possible to verify the presence of a peptide in a mixture. Furthermore, the identity of the parent protein could be determined based on the peptide fragmentation and sequence information provided by MSAD.

Peptide mixtures were also prepared with varying concentrations for each of the individual peptides in order to examine sensitivity and dynamic range. The mixtures contained all six peptides, as before (synthetic 3,4,6,7, angiotensin-1, meth-enkephalin), but one peptide was at a 1:100 concentration to the other 5 peptides. This 1:100 mixture ratio was repeated for each of the six peptides.

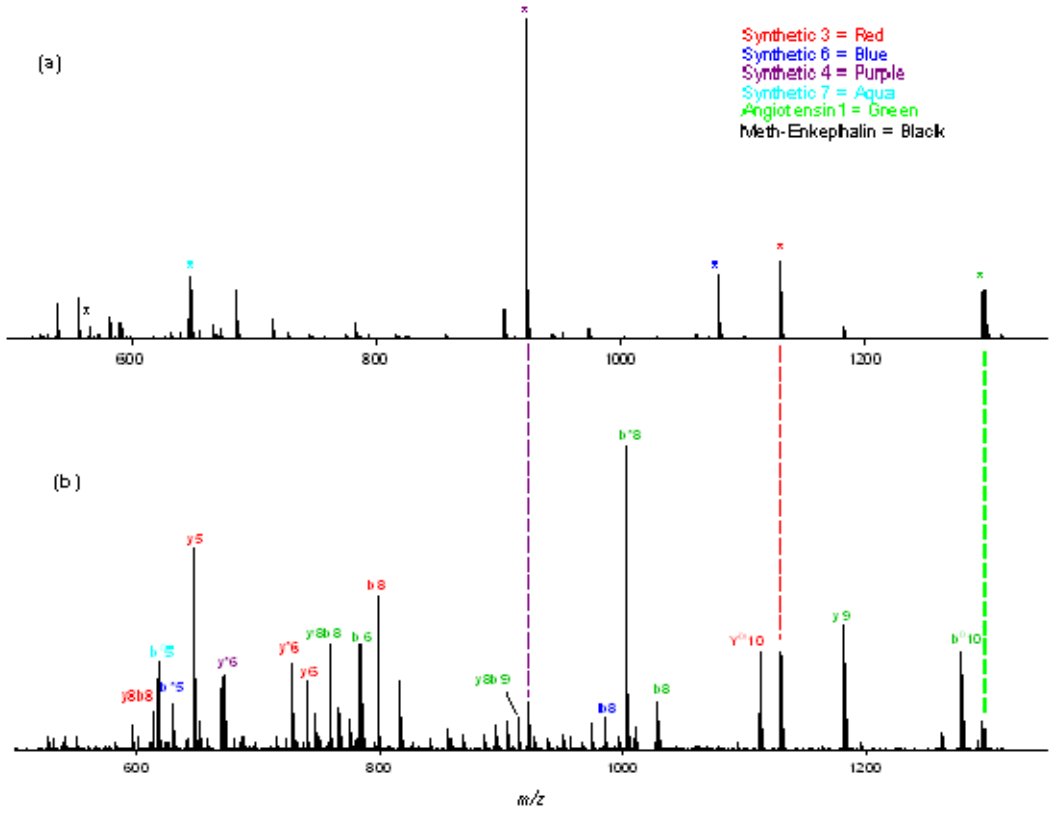


Figure 3.3: Dissociation data for 1:1 peptide mixture. A) FT-ICR spectrum of peptide mixture with each peptide mass peak labeled a different color. B) MSAD spectrum of peptide mixture with fragment ions labeled in the color corresponding to parent peptides. Dashed lines show the remaining parent masses in the spectrum.

Each parent peptide was assigned the same color label as in the 1:1:1:1:1:1 mixture, which was then used to label corresponding fragment ions within the MSAD spectra (Figure 3.4). Generally, the peptide that was at a lower concentration within the mixture produced MSAD fragments within the spectrum. This can be seen for the peptide mixture with angiotensin-1 at a lower concentration to the other five peptides (Figure 3.4). The major MSAD fragment ions are labeled within the spectrum to show that fragment ions from all six peptides are present (Figure 3.4). Only the most abundant fragment ions are labeled within the MSAD spectrum, although, the lower abundant fragment ions also provided identifications (Figure 3.4). For synthetic peptides 4 and 3, the parent ion remains within the spectrum (labeled with a (*) in the corresponding color). This identification of low abundance fragment ions demonstrates the good sensitivity and dynamic range of the MSAD method on peptide mixtures of varying concentration. The success rate of MSAD for providing fragment ions for all peptides in the mixture is 100%. To test the success rate of smaller peptides at lower concentrations meth-enkephalin was mixed at a lower concentration (1:100) to the other five peptides (data not shown). Again, there were identifiable fragment ions from all six peptides within the mixture. The fragment ions from meth-enkephalin were at a lower abundance within the spectrum, but still provide an identifiable isotopic packet when the peaks are expanded. In this study, the sensitivity and dynamic range afforded by MSAD is comparable to other methods of dissociation that do not require pre-isolation of parent ions.

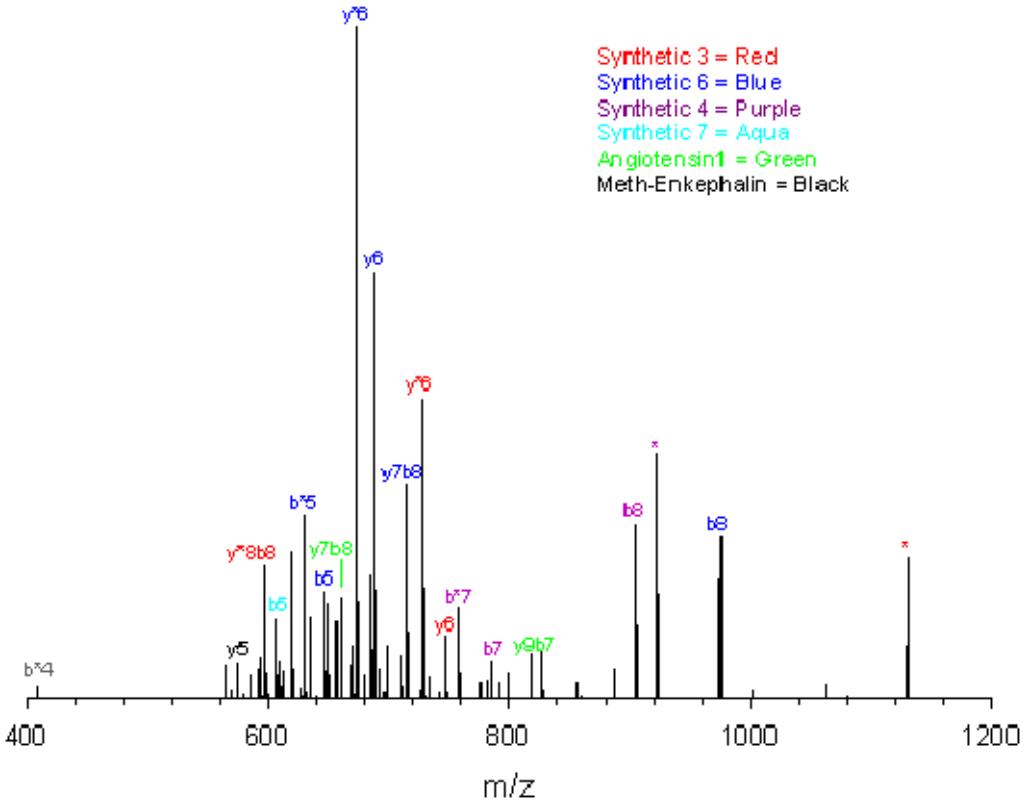


Figure 3.4: MSAD spectrum for six peptide mixture containing synthetic 3,4,6,7, angiotensin-1, and meth-enkephalin with angiotensin-1 at a 1:100 concentration to the other five peptides. The major fragment ions are color labeled, according to which parent peptide they were generated from, within the MSAD spectrum. Remaining parent peptides within the MSAD spectrum are labeled with a (*) and corresponding color.

MSAD forTryptic Digest of Apomyoglobin and BSA

Previously, only limited research has been conducted on MSAD of single peptides with no information being provided for more complex mixtures, especially tryptic digests. In general, most bottom-up proteomics work requires tryptic digest of intact proteins followed by some form of MS/MS analysis. MSAD allows for a MS/MS experiment on tryptic digest without pre-selection of parent ions, making it a useful sequencing tool that has a lower duty cycle with a smaller time scale than methods such as SORI-CAD. In this study, we have conducted MSAD experiments on tryptic digest of two large proteins, BSA and Apomyoglobin, to test the utility of MSAD for more complex peptide mixtures.

To examine the efficacy of MSAD on tryptic digests, a mass spectrum was obtained for the tryptic digest of apomyoglobin, followed by a MSAD spectrum (Figure 3.5a-b). The MSAD spectrum for the tryptic digest of Apomyoglobin shows 22 un-fragmented parent tryptic peptides (Figure 3.5b). There are 22 un-fragmented parent tryptic peptides remaining in the spectra, but most of these are lower in abundance. The five most abundant remaining parent ions within the apomyoglobin tryptic digest MSAD spectrum are labeled (*), revealing a wealth of abundant identifiable MSAD fragment ions (Figure 3.5b). Also, the MSAD spectrum of apomyoglobin contains a large number of internal fragment ions, as well as fragment ions that have a loss of water or ammonia, as seen in the single peptide and simple mixtures MSAD spectra (Table 3.4). This can be seen for identified apomyoglobin tryptic peptide 17-VEADIAGHGQEVLIR-31, spanning amino acids 17 through 31, which has all internal fragment ions with loss of water or ammonia (Table 3.4).

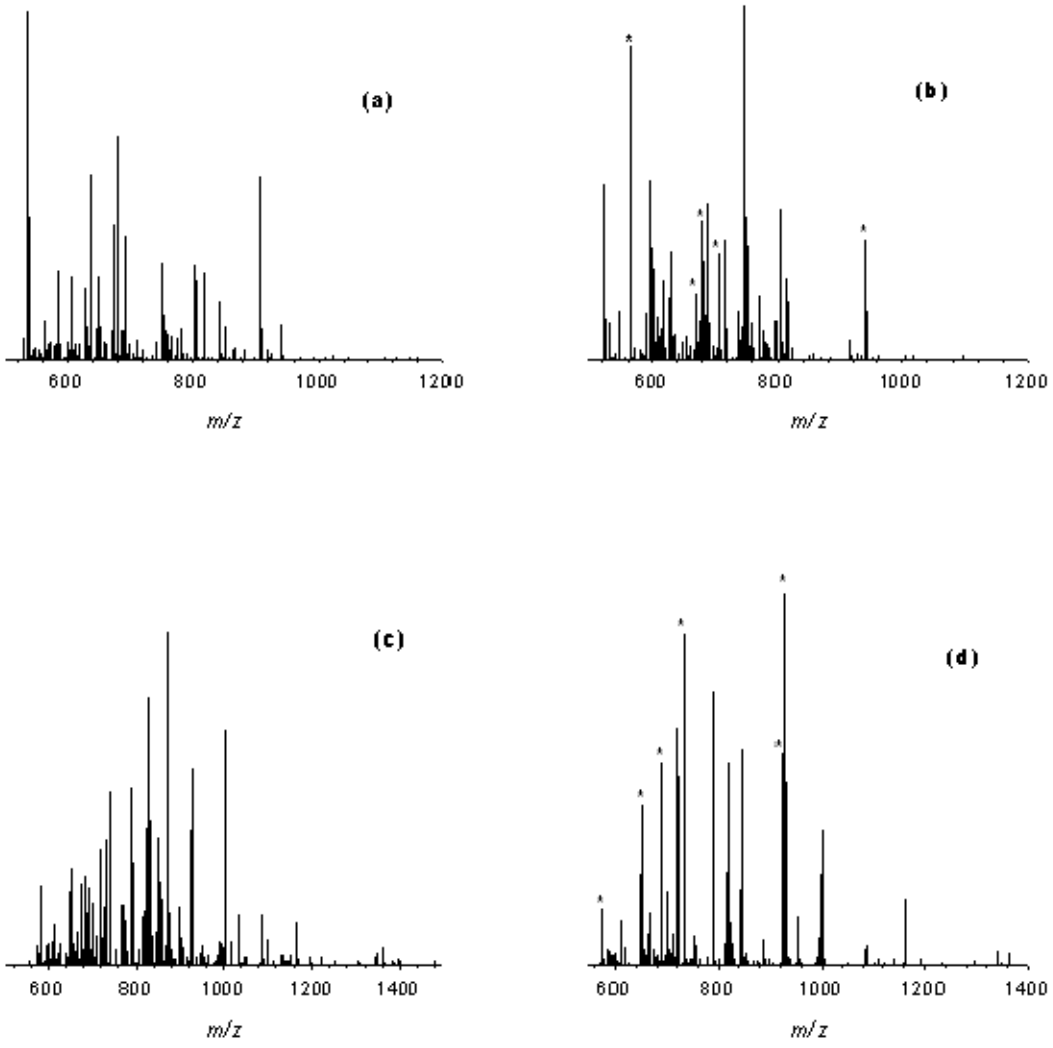


Figure 3.5: BSA and Apomyoglobin Tryptic digest MSAD spectrum. A) Apomyoglobin tryptic digest FT-ICR spectrum. B) Apomyoglobin tryptic digest MSAD spectrum with surviving parent masses labeled with (*). C) BSA tryptic digest FT-ICR spectrum. D) BSA tryptic digest MSAD spectrum with surviving parent masses labeled with (*).

Table 3.4: Apomyoglobin tryptic digest MSAD fragmentation data

Sequence of Tryptic Peptide	MSAD fragment ion (measured)	MSAD fragment ion (calculated)		
	Mass	Mass	ID	Sequence
32-LFTGHPETLEK-42	618.3272	618.32246	y5	ETLEK
	540.2588	540.25706	y10b6	FTGHP
32-LFTGHPETLEKFDKFKHLK-50	540.2588	540.25706	y18b6	FTGHP
	760.4654	760.40338	Y*9b*16	KFDKFK
	602.3585	602.31899	y*12b12	TLEKF
	618.3272	618.26494	y*16b*9	GHPETL
	585.335	585.29244	y*12b*12	TLEKF
32-LFTGHPETLEKFDKFKHLKTEAEMK-56	618.3272	618.26494	y*22b*9	GHPETL
	585.335	585.29244	y*18b*12	TLEKF
	602.3585	602.31899	y*18b12	TLEKF
	737.4197	737.431	y10b21	KHLKTE
32-LFTGHPETLEKFDKFK-47	618.3272	618.28875	y*13b9	GHPETL
	540.2588	540.25706	y18b6	FTGHP
	585.335	585.29244	y*9b*12	TLEKF
	760.4654	760.40338	y*6b*16	KFDKFK
1-GLSDGEWQQVLNVWGKVEADIAGHGQEVLIR-31	540.2588	540.23057	y*14b23	EADIAG
	687.3822	687.31021	y*9b*29	GHGQEVL
	660.3053	660.26293	y*14b*24	EADIAGH
	706.4031	706.35643	y*23b*14	QVLNVW
	679.4	679.33028	y*17b*21	GKVEADI
	588.2961	588.2418	y*13b*25	ADIAGHG
119-HPGDFGADAQGAMTK-133	626.2714	626.20983	y*14b*8	PGDFGAD
103-YLEFISDAIIHVLHSKHPGDFGADAQGAMTK-133	626.2714	626.20983	y*14b*24	PGDFGAD
	588.2961	588.2418	y*17b*20	ADIAGHG
	719.4053	719.37281	y*26b12	SDAIIHV
	613.3476	613.3561	y27b10	ISDAII
97-HKIPIKYLEFISDAIIHVLHSKHPGDFGADAQGAMTK-133	588.2961	588.2418	y*17b*26	SKHPGD
	719.4053	719.37281	y*26b18	SDAIIHV
	579.3518	579.303	y*27b*16	ISDAII
	630.3732	630.35029	y*33b9	IKYLE
97-HKIPIKYLEFISDAIIHVLHSK-118	719.4053	719.37281	y*11b18	SDAIIHV
	613.3476	613.3561	y12b16	ISDAII
	579.3518	579.303	y*12b*16	ISDAII
	630.3732	630.35029	y*18b9	IKYLE
103-YLEFISDAIIHVLHSK-118	719.4053	719.40786	y6	HVLHSK
	613.3476	613.3561	y12b10	ISDAII
	579.3518	579.303	y*12b*10	ISDAII
17-VEADIAGHGQEVLIR-31	687.3822	687.31021	y*9b*13	GHGQEVL
	660.3053	660.26293	y*14b*8	EADIAGH
	588.2961	588.2418	y*13b*9	ADIAGHG

Table 3.4: Continued

	MSAD fragment ion (measured)	MSAD fragment ion (calculated)		
Sequence of Tryptic Peptide	Mass	Mass	ID	Sequence
1-GLSDGEWQQVLNVWGK-16	706.4031	706.35643	y*8b*14	QVLNVW
	585.335	585.31491	y5b16	NVWGK
78-KKGHHEAELKPLAQSHATK-96	660.3053	660.32178	y18b7	KGHHEA
	803.4717	803.40518	y*8b*19	LAQSHATK
	708.3645	708.36806	y15b10	HEAELK
	692.365	692.33676	y*8b16	LAQSHAT
	617.3262	617.30474	y*9b16	PLAQSH
	537.3076	537.25605	y*14b*10	EAELK
79-KGHHEAELKPLAQSHATK-96	803.4717	803.40518	y*8b*18	LAQSHATK
	708.3645	708.36806	y15b9	HEAELK
	692.365	692.33676	y*8b17	LAQSHAT
	617.3262	617.30474	y*9b15	PLAQSH
	537.3076	537.25605	y*14b*9	EAELK

There are also many identified apomyoglogin peptides in the mixture that have missed enzymatic cleavages. This phenomenon is evident when examining the identified sequences of the fragment ions. For example, the first four peptides in Table 3.4 all start at amino acid number 32 but end at varying lengths from amino acid number 42 to 56 (Table 3.4). These missed cleavages create a series of fragment ions that have the same mass and sequence for the different tryptic peptides. However, there are repeating fragments within the MSAD spectrum due to missed enzymatic cleavages, the fragment ions obtained allow for the identification of these missed cleavage locations within the amino acid sequence of the peptide and protein. Therefore, the extensive fragmentation provided by MSAD gives the ability to identify peptides with missed enzymatic cleavages.

In order to examine efficacy of using MSAD on a tryptic digests of a large protein, tryptic digests of BSA were used in this study. For comparison, a mass spectrum was obtained for the tryptic digest of BSA followed by a MSAD spectrum (Figure 3.5c-d). Similar to apomyoglobin, the MSAD spectrum of the BSA tryptic digest reveals 20 un-fragmented parent tryptic peptides within the spectrum (Figure 3.5d). Of these 20 un-fragmented parent peptides, only six (labeled with a (*)) are abundant within the MSAD spectrum (Figure 3.5d). Again, similar to apomyoglobin, the MSAD spectrum of BSA tryptic digest contains a large number of internal fragment ions with the loss of water and/or ammonia (Table 3.5). The fragment ions obtained by MSAD for BSA are more distinct, with less repeating fragment ions, than apomyoglobin. However, there are still identified missed enzymatic cleavages for BSA as well as identified non-tryptic peptides.

Table 3.5: BSA tryptic digest MSAD fragmentation data

	MSAD fragment ion (measured)	MSAD fragment ion (calculated)		
Sequence of Tryptic Peptide	Mass	Mass	ID	Sequence
224-LSQKFPK-230	588.3166	588.31457	y6b6	SQKFP
143-YLYEIAR-149	616.3291	616.30949	y*5b7	YEIAR
580-LVVSTQTALA-589	616.3291	616.33062	y9b7	VVSTQT
506-AFDEKLFTFHADICTLPDTEK-526	616.3291	616.29825	y*20b6	FDEKL
	616.3291	616.29825	y*19b7	DEKLF
	751.3986	751.3986	y*7b*21	TLPDTEK
511-LFTFHADICTLPDTEKQIK-529	751.3986	751.3986	y*10b*16	TLPDTEK
	716.365	716.3675	b6	LFTFHA
268-YICNQDTISSKLK-281	711.4109	711.35649	y*8b*13	DTISSKL
384-HLVDEPQNLIKQNCDQFEK-402	711.4109	711.41535	y12b13	NLIKQN
	810.4057	810.39976	y8b10	DEPQNLI
231-AEFVEVTKLVTDTKVHKECCHGDLLE	711.4109	711.35649	y*13b*50	DTISSKL
CADDTADLAKYICDNQDTISSKLKECCDK-286	887.5088	887.52021	y12b52	TISSKLKE
384-HLVDEPQNLIK-394	810.4057	810.39976	y8b10	DEPQNLI
329-DAFLGSFLYEYSR-341	810.4057	810.40378	y10b10	LGSFLYE
551-TVMENFVAFVDK-562	660.3739	660.37209	y6b12	VAFVDK
27-GLVLIAFSQYLQQCPFDEHVK-47	660.3739	660.37209	y18b9	LIAFSQ
342-RHPEYAVSVLLR-353	663.3358	663.27785	y*11b*7	HPEYAV
83-VASLRETYGDMADCCEKQE	663.3358	663.3466	y35b8	LRETY
PERNECFLSHKDDSPDLPK-120	751.3986	751.3986	y*22b22	KQEPER
	716.365	716.36192	y14b30	CFLSHK
83- VASLRETYGDMADCCEKQEPERNECFLSHKD	663.3358	663.3466	y57b8	LRETY
DSPDLPKLKPDPNTLCDEFKADEKKFWGK-142	751.3986	751.3986	y*44b22	KQEPER
	716.365	716.36192	y36b30	CFLSHK
	887.5088	887.47795	y*7b60	EKKFWGK
	647.4093	647.37684	y*27b39	PDLPKL
292-SHCIAEVEKDAIPENLPPLTADFAEDK	751.3986	751.3986	y*28b*10	IAEVEKD
DVCK-322	716.365	716.36192	y14b24	PLTADFA
246-VHKECCHGDLLECADDRADLAK-267	655.3993	655.38913	y6b22	RADLAK
121-LKPDPNTLCDEFKADEKKFWGK-142	887.5088	887.47795	y*7b22	EKKFWGK

For example, peptide 224-LSQKFPK-230 has a missed cleavage, but the most abundant fragment ion y6b6 was able to provide an identification of the peptide as well as the verification of a missed cleavage site (Table 3.5). Also identified was the non-tryptic BSA peptide 580-LVVSTQTALA-589 from the predominant fragment ion y9b7 (Table 3.5). The ability to identify missed cleavages and non-tryptic peptides provides another argument for the viability of MSAD with complex peptide mixtures.

Identification of MSAD fragments from BSA and Apomyoglobin were made using the PROWL website mass spectrometry fragmentation function. Fragment ion identification in PROWL is output with two decimal places. In order to obtain a more accurate mass match for each PROWL identified MSAD fragment ion, the fragment ion mass was calculated to four decimal places in order to match back to the mass of the MSAD fragment from the spectrum. Identification of MSAD fragments from BSA and Apomyoglobin using PROWL and calculated masses show several possible tryptic fragments corresponding to each identifiable MSAD fragment in the spectra (Tables 3.4 & 3.5). However, this is in part a consequence of multiple tryptic peptides that are capable of producing fragment ions with the same sequence. There is also a preference toward certain fragment ions within the MSAD spectrum. Each tryptic peptide gives a preferential MSAD fragment ion that can have different combinations of water and ammonia loss (Table 3.4 & 3.5).

The use of MSAD, as a replacement for more commonly applied fragmentation methods such as SORI-CAD with a FT-ICR is a feasible option for simple peptide solutions, tryptic digest and simple mixtures. MSAD provides a fragmentation method that can fragment all peptides in the sample, in one step, eliminating the isolation step

needed for SORI-CAD, which provides a more operationally simple and time saving method. This is especially important if sample limitation is of concern. On preliminary inspection, the MSAD method provides a more extensive identifiable fragmentation pattern than SORI-CAD. However, the MSAD method does lead to more internal fragment ions making identification more complicated. The MSAD method works well on simple peptides, but when applied to complex tryptic digest the lack of fragmentation of major parent ions and the number of internal fragment ions produced does present a rather complex spectrum, but this is not of enough significance not to provide identification of the peptide or protein. A large number of MSAD fragments from tryptic peptides were identified. MSAD on simple mixtures gives a rather rich spectrum of identifiable fragment ions when the peptides are at equal concentrations. The sensitivity of MSAD could provide some problems, but for complex mixtures being examined by FT-ICR-MS where rapid dissociation of parents ions is needed MSAD provides a very useful alternative to SORI-CAD.

Conclusions

Through these two studies, better methods for protein charge state determination under liquid chromatography conditions and fragmentation of proteins and peptides within the FTICR-MS were examined. By applying new methods, such as the TACT program and MSAD fragmentation, fundamental advancements in these areas were made. The TACT program allowed for the determination of large proteins charge states under liquid chromatography time frames better than previously applied software. Also shown in this study, was that complex mixtures being examined by FT-ICR-MS, where rapid

dissociation of parent ions is needed, MSAD provides a very useful alternative to SORI-CAD.

Chapter 4

Application of the Integrated Top-Down Bottom-Up Methodology for the Characterization of Ribosomal Protein Mixtures for PTMs and Isoforms

Data presented below is in final preparation for submission or published as the following

Heather M. Connelly, Eric Hamlett, David Robinette, Kevin Ramkissoon, Hsun-Cheng Hsu, Ming Yu, Robert L. Hettich, and Morgan C. Giddings. Characterization and Comparison of Ribosomal Protein Heterogeneity and Isoforms in Wild-Type and Variant Strains of *E. coli*. *Nature Biotechnology, In final preparation* (2006). *All FTICR top-down and LCQ Bottom-up sample preparation, experiments and data analysis were performed by Heather M. Connelly.*

Strader, M.B.; VerBerkmoes, N.C.; Tabb, D.L.; Connelly, H.M.; Barton, J.W.; Bruce, B.D.; Pelletier, D.A.; Davison, B.H.; Hettich, R.L.; Larimer, F.W.; and G.B. Hurst. Characterization of the 70S Ribosome from *Rhodopseudomonas palustris* using an Integrated "Top-Down" and "Bottom-Up" Mass Spectrometric Approach. *Journal of Proteome Research,* 2004; 3, 965-978. *All bottom-up MS, sample preparation, experiments and data analysis on Rhodopseudomonas ribosomal complex were performed as a joint effort between Nathan C. VerBerkmoes, Brad Strader,and David Tabb, All top-down experiments and data analysis was performed by Heather M. Connelly with assistance from Robert L. Hettich.*

Introduction

Integrating "top-down" and "bottom-up" MS-based proteomic strategies provides a powerful tool to examine complex protein mixtures, such as proteins in multi-component complexes, or even complete proteomes. The first of these methods is intact protein, or top-down mass spectrometry, which can be used to provide intact protein identification, as well as insight into protein modification states. This powerful method can provide information on the natural state of intact proteins, including details about post-translational modifications (PTM's), truncations, mutations, signal peptides, and isoforms, due to the ability to measure the molecular mass of a protein very accurately

and detect any covalent modifications that alter the mass of a protein. The top-down mass spectrometry approach for proteins was first introduced with electrospray-Fourier transform ion cyclotron resonance mass spectrometry (ESI-FTICR-MS) [22, 23, 24]. The dynamic range, sensitivity, and mass accuracy offered by high performance FT-ICR-MS affords not only unambiguous protein identification in many cases, but also detailed information about protein modifications. Eventough, top-down methodologies provide a powerful analytical approach, some limitations do exist; on-line chromatography of intact proteins is often difficult due to the wide range of protein sizes and hydrophobicities. Furthermore, data are often difficult to analyze and interpret due to limited bioinformatics tools.

The more common peptide or "bottom-up" mass spectrometric approach involves enzymatic digestion of intact proteins with a protease such as trypsin, Glu-C or cyanogen bromide in order to generate a peptide mixture. This peptide mixture is then analyzed by MS/MS methods to generate peptide fragmentation spectra that are compared back to a database with searching algorithms. This "bottom-up" proteomics approach is able to quickly and efficiently provide a comprehensive list of proteins present in a large multi-protein complex. Bottom-up methods provide a comprehensive list of proteins, although, vital information about post translational modifications may be missed if the peptides containing the particular modification escape detection. Furthermore, identifying peptides that come from a complex protein mixture does not provide information on the presence of different isoforms that may exist for a particular protein.

An integrated top-down and bottom-up approach allows for a more comprehensive characterization of protein complexes due to the unique strength of each

technique. In an integrated approach, intact protein masses from the top-down analysis corresponding to a particular PTM or isoform can be compared to the comprehensive list of proteins provided by the bottom-up analysis. This correlation between the two methods can provide PTM location and identity with more certainty. The comprehensiveness of this technique has been previously demonstrated in a study of the *Shewanella oneidensis* proteome [35].

Since this technology was to be ultimately used for whole proteomes under multiple growth states (Chapter 7), we started with the 70S Ribosome from *Rhodopseudomonas palustris* and progressed our technique into the examination of ribosomal proteins from multiple strains of antibiotic resistant *E. coli*. The ribosome has been a model protein complex for the development of MS-based proteomics techniques; due to the ease of purification, the limited complexity and the presence of numerous post-translational modifications [88]. The ribosome is the universal macromolecular machine involved in translating the genetic code into proteins. Bacterial ribosomes are composed of a small subunit (30S) containing about 20 proteins and a single rRNA (16S), and a large subunit (50S) consisting of over 30 proteins and two rRNAs (23S and 5S). The bacterial ribosomal proteins have been shown to be well conserved across different species, and this includes their PTMs.

In our first study, the ribosomal proteins from *R. plaustris* were examined for positive identification of the protein, as well as identification of associated PTMs. For this study, the bottom-up approach was expanded to the use of 1D and 2D LC-MS/MS methodologies for the analysis of the enzymatically digested protein complex. This was necessary due to the increased complexity of the protein complex. The top-down

methodology was performed with the high resolution and high mass accuracy FT-ICR instrument. For these experiments, we performed LC-ES-FT-ICR for intact protein measurements. Using this integrated approach, we were able to identify a complement of ribosomal proteins and their associated PTMs.

In the second study, we employ an integrated top-down and bottom-up approach to characterize the ribosomal proteins from wild type K12 and two streptomycin resistant strains of *E. coli*. Using this method, a complement of ribosomal proteins with unique PTM series, isoforms, and point mutations were identified from all three strains. With this integrated top-down and bottom-up approach, we were able to provide a more comprehensive examination of the role of ribosomal proteins in antibiotic resistance than if an individual method had been employed.

Results Characterization of the *R. palustris* 70S Ribosome

Top-down and Bottom-up Characterization of the R. palustris 70S Ribosome

The 70S ribosome from *R. palustris* was characterized with the integrated top-down and bottom-up technique. Integration of results was achieved by using protein identifications from the analysis of top-down data to refine analysis of bottom-up data, and vice versa, in an iterative manner to increase the number of characterizations of ribosomal proteins obtained. For example, identification of a methylated protein by the top-down approach could provide motivation to examine more closely the bottom-up results for the presence of a methylated peptide from that protein. The combined top-down bottom-up MS analysis identified a total of 53 of the predicted 54 ribosomal proteins [Table 4.1]. The data indicated the presence of 21 proteins for the small subunit and 33 for the large subunit (S20 and L26 are identical). No orthologue of *E. coli* S22

was identified for *R. palustris* ribosomes. We also identified isoforms for L7/L12 from the large subunit. These isoforms included one form with 3 methylations and a second form with an acetylation. Within this work, each of the *R. palustris* ribosomal proteins (RRP) is named after the corresponding ribosomal protein in *E. coli*. The L7/L12 isoforms were therefore named RRP-L7/L12A and RRP-L7/L12B.

Intact proteins from three separate aerobically grown ribosome samples were examined by LC-FT-ICR-MS, and the resulting data were pooled. From this top-down analysis, we identified 42 intact *R. palustris* ribosomal proteins. The four largest ribosomal proteins (RRP-S2 at 36 kDa, RRP-S1 at 62.8 kDa, RRP-L2 at 31.6 kDa, and RRP-S3 at 26.3 kDa) were not observed. Even though the FT-ICR-MS has sufficient mass range to observe these species, prior experience with intact proteins suggests that larger species, such as these, are difficult to elute from the C4 reverse-phase column under the experimental conditions employed for the top-down liquid chromatography. It is likely that the increased hydrophobicity of these larger proteins results in irreversible binding on the reverse-phase column, making these proteins difficult, if not impossible, to elute from the column.

Table 4.1. Ribosomal protein identification by top-down ESI-FTICR-MS [54]

Protein	Modification	Calc. Mass[a]	Meas. Mass	Mass error (ppm)
L1	loss of Met	23877.832	23877.449	*16.0*
L3	plus Methyl	25622.463	25622.159	*11.9*
L5	plus 2 Methyl	21064.992	21064.576	*19.7*
L6	loss of Met	19272.408	19272.674	*-13.8*
L7/L12	loss of Met + 3 Methyl	12754.07	12754.089	*-1.5*
L9	none	21178.022	21178.268	*-11.6*
L10	loss of Met	19067.739	19067.617	*6.4*
L11	loss of Met+Acet+ 9 Methyl	15507.107	15507.246	*-9.0*
L14	none	13488.498	13488.645	*-10.9*
L15	none	16836.243	16836.259	*-1.0*
L17	plus 3 Methyl	15716.353	15716.056	*18.9*
L18	loss of Met	12904.93	12905.157	*-17.6*
L19	none	14296.764	14296.899	*-9.4*
L21	loss of Met	13358.081	13358.533	*-33.8*
L22	loss of Met	13826.007	13825.6447	*26.2*
L23	none	10907.949	10908.021	*-6.6*
L24	loss of Met	10998.226	10998.231	*-0.5*
L24	loss of Met + Methyl	11012.241	11012.146	*8.6*
L29	loss of Met	7849.213	7849.239	*-3.3*
L30	loss of Met	7092.967	7092.988	*-3.0*
L31	none	8566.315	8566.334	*-2.2*
L32	loss of Met	6860.73	6860.636	*13.7*
L33	loss of Met + Methyl	6248.504	6248.45	*8.6*
L35	loss of Met	7415.278	7415.278	*0.0*
L36	none	5063.971	5063.952	*3.8*
S4	loss of Met + Methyl	23441.536	23441.69	*-6.6*
S5	loss of Met	20522.086	20522.411	*-15.8*
S7	loss of Met	17556.27	17556.629	*-20.4*
S8	loss of Met	14477.6316	14477.683	*-3.6*
S8	loss of Met+Acet+4 Methyl	14575.704	14575.619	*5.8*
S10	none	11667.363	11667.404	*-3.5*
S11	loss of Met + Methyl	13760.215	13760.314	*-7.2*
S12	none	13874.799	13875.167	*-26.5*
S13	loss of Met	14313.985	14313.596	*27.2*
S14	loss of Met	11331.399	11331.9	*-44.2*
S15	loss of Met	10010.563	10010.562	*0.1*
S16	loss of Met	12017.595	12017.575	*1.7*
S17	loss of Met	9553.253	9553.316	*-6.6*
S18	plus 6 Methyl	9178.219	9177.834	*41.9*
S19	loss of Met	10087.371	10087.379	*-0.8*
S20	loss of Met	9577.324	9577.387	*-6.6*
S21	none	10062.669	10062.722	*-5.3*

[a] MAIM (most abundant isotopic mass)

In total, 42 proteins were tentatively identified, with the majority (25) at better than 10 ppm mass accuracy, and only 3 differing by >30 ppm from the calculated value. Of these 42, ten correspond directly to the predicted gene products, 21 are processed by only methionine truncation, and the remaining 11 appear to be modified by further acetylation and/or methylation. Three proteins, RRP-L24, RRP-L7/L12 and RRP-S8, were found to be present in two different forms. The most highly modified species identified was RRP-L11, which is methionine-truncated, and contains multiple methylations and/or acetylations. About ten additional species were measured from the ribosome sample, but could not be identified. It is likely that these species correspond to the other ribosomal proteins, but are altered substantially (possibly by combinations of other PTMs, oxidation, and more extensive truncation) such that they are beyond the scope of our simple "look-up table" (excel table with all combinations of searched for PTMs) or they could be common contaminants identified in the bottom-up analysis as well. Using this integrated approach for the *R. palustris* ribosomal proteins we were able to provide a comprehensive analysis of PTMs and isoforms that was previously unknown for this organism.

Results for *E. coli* Ribosomal Proteins From All Three Strains

General Analysis of E. coli Ribosomal Proteins from All Three Strains

Proteins from three strains of K12 *E. coli* were examined with a combined top-down and bottom-up strategy. The three strains included a K12 wild type strain (WT), a K12 streptomycin resistant strain (SmR), and a K12 streptomycin resistant compensated strain where cell growth was allowed to return to a normal state (SmRC). To obtain the accurate mass values for the top-down measurements, the most abundant isotope

measurement method (MAIM) was used as previously described [89]. In this method the MAIM value are obtained for the top-down measured masses and then compared to the calculated MAIM values for each ribosomal protein. Bottom-up identifications are also made for each ribosomal protein with the number of unique peptides identified and protein sequence coverage recorded. To investigate the fidelity of the top-down database searching, two distracter database searches were performed with the bacterium *Rhodopseudomonas palustris* and yeast *Saccharomyces cerevisiae* ribosomal protein databases plus the entire *E. coli* database to see how many proteins are identified using the measured *E. coli* protein masses. When using the *R. palustris* ribosomal database, five *E. coli* ribosomal measured masses are identified within 1.0 Da from the *R. palustris* database; these include L31, S17, S10, L36, and L28. L31 was identified with to have a N-terminal methionine truncation for *E. coli* that was not identified in *R. palustris* search. For the searches against the yeast database, only three yeast proteins were identified within 1 Da using the measured *E. coli* masses, including the 60S L28, 60S L44, and 40S S21 proteins. The yeast 60S L44 protein has homology to the *E. coli* L12 protein which could provide a match within the yeast database.

In the WT strain measurement, a total of 52 of the 57 ribosomal proteins were identified by the bottom-up approach and 43 of the 57 were identified by top-down analysis [Table 4.2]. The bottom-up analysis of the WT strain indicated the presence of 20 out of a total 22 proteins from the small subunit (denoted S1-S22), with S12 and S22 not being seen, and 30 of the possible 36 proteins from the large subunit (denoted L1-L36), with L26, L31, and L34-L36 not being identified [Table 4.2].

Table 4.2: Combined top-down and bottom-up data for the WT strain.

Subunit	Avrg Sequence Mass	Measured Mass	PTM	PPM	BU Seq Cov	Unique Peptides
S1					53.9	24
S2					54.8	9
S3	25852.07850	25852.80623	DEM	-28.14992226	44.2	13
S4	23337.93550	23337.98857	DEM	-2.27402291	37.9	9
S5	17514.26790	17514.22458	DEM, ACE	2.473640363	62.3	12
S6					34.4	3
S7	19887.93140		DEM		45.8	13
S8	13995.39720	13995.439	DEM	-2.986624774	27.7	3
S9	14725.03120		DEM		13.8	2
S10	11735.60360	11735.4623		12.04053961	47.6	6
S11	13727.78420		DEM, MET		39.5	3
S12	13651.88700	13652.15343	DEM, BMT	-19.51598339		
S13					53.4	10
S14	11449.31400	11448.96907	DEM	30.12687048	12.9	1
S15	10137.58300		DEM		7.9	2
S16	9190.56590	9190.601945		-3.921956536	30.5	3
S17	9573.27380	9573.259505	DEM	1.493219592	9.5	1
S18					29.3	2
S19	10299.11000	10298.42837	DEM	66.18348576	12	3
S20	9553.21800	9553.368212	DEM	-15.72370692	12.6	1
S21	8368.77370	8368.72685	DEM	5.598191764	14.1	1
S22						
L1	24598.48930	24598.41311	DEM	3.097547946	59	16
L2	28729.30750	28728.46272	DEM	29.40499001	34.4	9
L3	22257.57560	22257.41732	MET	7.111421425	33.5	8
L4					28.4	4
L5	20170.42370	20170.53916	DEM	-5.724074106	57.5	12
L6	18772.61160		DEM		49.9	10
L7	12206.06290	12206.05081	DEM, ACE	0.990737153	64.5	10
L9					51	9
L10	17580.43760	17579.97320	DEM	26.41566783	31.5	5
L11	14870.47030	14870.38882	DEM, 9-MET	5.479315607	26.1	6
L12	12206.06290	12206.05081	DEM, ACE	0.990737153	64.5	10
L13	16918.57380	16918.02986		32.15052323	34.5	4
L14	13541.06560	13540.54586		38.38242981	21.1	3
L15	14980.44430	14980.42223		1.472987019	34.7	5
L16					32.4	4
L17	14364.62170	14364.05432		39.49877775	26.8	5
L18	12769.64490	12769.87237		-17.81365118	19.7	2
L19	13002.05480	13001.53381	DEM	40.06982035	36.5	4
L20	13365.77070		DEM		28	6
L21	11565.05541	11564.3661		59.60239496	44.7	5
L22					42.7	7
L23	11199.13960				27	3
L24	11185.02740	11185.06116	DEM	-3.017873698	52.9	8
L25	10693.46300	10693.44982		1.23280924	60.6	8
L26						
L27	8993.28970	8993.038869	DEM	27.89090626	25.9	2
L28	8875.31060	8874.791879	DEM	58.4453912	12.8	1
L29	7273.46450	7273.217968		33.89471414	46	3
L30	6410.61260	6410.67098	DEM	-9.106773977	33.9	2
L31	7871.10070	7871.263986		-20.74500203		
L32	6315.19890	6314.680172	DEM	82.139614	26.3	1
L33	6254.42280	6254.571417	DEM, MET	-23.76190494	27.3	1

DEM = N-terminal methionine truncation
ACE = acetylation
MET = Methylation
BMT = beta- methylthiolation and a K to T point mutation

In the top-down analysis 10 of the small subunit proteins were not identified including S1, S2, S6, S7, S9, S11, S13, S15, S18, and S22; while 11 proteins from the large subunit were not found including L4, L9, L6, L16, L20, L22, L23, L26, and L34-L36 [Table 4.2]. Ribosomal proteins from the WT strain, not found by bottom-up and top-down, include S22, L26, and L34-L36.

From the SmR strain analyses 44 of the 57 ribosomal proteins were identified by the bottom-up method while 41 were identified using the top-down approach [Table 4.3]. The data from the SmR strain shows the S13 and S22 proteins from the small subunit as well as the of L26, L27, L30-L32, and L34-L36 proteins from the large subunit were not detected by bottom-up analysis [Table 4.3]. In the SmR strain the S17, S22, L26, L31, and L34-L36 were not found by bottom-up and top-down measurements.

Within the SmRC strain of streptomycin resistant *E. coli* 49 of the 57 ribosomal proteins were identified by bottom-up methods and 43 by the top-down method [Table 4.4]. From the small subunit ribosomal proteins of the SmRC strain the S22 protein was not observed by bottom-up analysis; while the large subunit proteins L26, and L34-L36 were not observed [Table 4.4]. The top-down analysis shows the S1, S2, S6, S13, S18, and S22 proteins from the small subunit not detected and the L4, L9, L22, L26, L31, and L34-L36 proteins from the large subunit not detected [Table 4.4]. Top-down and bottom-up measurements did not identify the S22, L26, and L34-L36 ribosomal proteins from the SmRC strain.

Table 4.3: Combined top-down and bottom-up data for the SmR strain.

Subunit	Avrg Sequence Mass	Measured Mass	PTM	PPM	BU Seq Cov	Unique Peptides
S1					46.3	25
S2					54.8	14
S3	25852.07850	25852.1039	DEM	-0.982551558	42.1	13
S4	23337.93550	23337.94815	DEM	-0.542035948	40.3	12
S5	17514.26790	17514.30762	DEM, ACE	-2.267808179	62.9	14
S6					42.7	4
S7	19887.93140	19887.97468	DEM	-2.176194151	46.4	15
S8	13995.39720	13995.30806	DEM	6.369594141	21.5	2
S9	14725.03120	14724.9854	DEM	3.110349946	20	5
S10	11735.60360	11735.49733		9.055350165	13.6	1
S11	13727.78420	13727.72578	DEM, MET	4.255603027	40.3	6
S12	13651.88700	13624.85	DEM, BMT, K-T		21.8	1
S13					53.4	10
S14	11449.31400				12.9	1
S15	10137.58300	10137.89412	DEM	-30.68966242	33.7	1
S16	9190.56590	9190.203623		39.41835616	13.4	1
S17	9573.27380	9572.489042	DEM	81.97383846		
S18					13.3	1
S19	10299.11000	10299.02371	DEM	8.378393861	20.7	3
S20	9553.21800	9552.866332	DEM	36.81147023	24.1	4
S21	8368.77370	8368.720287	DEM	6.382416578		
S22						
L1	24598.48930	24598.32132	DEM	6.828752691	59	20
L2	28729.30750				37.4	12
L3	22257.57560	22257.60407	MET	-1.279249839	33.5	10
L4					28.4	7
L5	20170.42370	20170.28273	DEM	6.988797166	57.5	14
L6	18772.61160	18772.58227	DEM	1.562169432	40.7	11
L7	12206.06290	12206.52929	DEM, ACE	-38.20937216	64.5	9
		12220.1022	DEM, ACE, MET			
L9					36.2	9
L10	17580.43760	17580.72068	DEM	-16.10215891	31.5	9
L11	14870.47030	14870.58317	DEM, 9-MET	-7.590210513	19.7	3
L12	12206.06290	12206.52929	DEM, ACE	-38.20937216	64.5	9
L13	16918.57380	16917.8734		41.39840676	35.9	5
L14	13541.06560	13541.00182		4.709969059	28.5	4
L15	14980.44430	14980.63423		-12.6781954	34.7	8
L16					32.4	6
L17	14364.62170				26.8	5
L18	12769.64490	12770.35352		-55.49285086	19.7	2
L19	13002.05480	13002.40935	DEM	-27.2689206	36.5	4
L20	13365.77070	13365.48241	DEM	21.56905176	16.1	4
L21	11565.05541				44.7	3
L22	12226.31560	12226.03966		22.56935033	51.8	9
L23	11199.13960	11199.07846		5.459347966	15	1
L24	11185.02740	11185.04528	DEM	-1.598207976	44.2	8
L25	10693.46300	10693.45309		0.927108459	25.5	2
L26						
L27	8993.28970	8993.445437	DEM	-17.31702249		
L28	8875.31060				12.8	1
L29	7273.46450	7272.763565		96.36879372	22.2	1
L30	6410.61260	6410.312477	DEM	46.81658661		
L31	7871.10070	7871.051347		6.27015228		
L32	6315.19890	6314.651869	DEM	86.62134141		
L33	6254.42280	6254.318399	DEM, MET	16.69234769	27.3	1
L34						
L35						
L36	4364.35210	4364.816693		-106.4517686		

DEM = N-terminal methionine truncation
ACE = acetylation
MET = Methylation

Table 4.4: Combined top-down and bottom-up data for the SmRC strain.

Subunit	Avrg Sequence Mass	Measured Mass	PTM	PPM	BU Seq Cov
S1					52.8
S2					45.6
S3	25852.07850	25852.02755	DEM	1.970905357	42.1
S4	23337.93550	23337.41086	DEM	22.48026609	32.00
S5	17514.26790	17514.31651	DEM, ACE	-2.775565629	57.5
S6					34.4
S7	19887.93140	19887.90318	DEM	1.418850429	46.4
S8	13995.39720	13995.38833	DEM	0.633851249	34.6
S9	14725.03120	14724.25623	DEM	52.62949799	20
S10	11735.60360	11735.62495	DEM	-1.818824214	34
S11	13727.78420	13727.65999	DEM, MET	9.048364848	39.3
S12	13651.88700	13624.78221	DEM, BMT, K-T		21.8
S13					53.4
S14	11449.31400	11449.16009	DEM	13.44263945	12.9
S15	10137.58300	10137.58642	DEM	-0.336865306	6.7
S16	9190.56590	9190.203623		39.41835616	30.5
S17	9573.27380	9573.121732	DEM	15.88463917	9.5
S18					29.3
S19	10299.11000	10298.68954	DEM	40.8251781	29.3
S20	9553.21800	9553.340992	DEM	-12.87440525	12.6
S21	8368.77370	8368.84807	DEM	-8.886606648	14.1
S21	8466.96200	8466.674	7 MET	34.01456154	
S22					
L1	24598.48930	24598.81402	DEM	-13.20072936	59
L2	28729.30750	29729.05438	DEM	-34798.85058	34.4
L3	22257.57560	22257.65787	MET	-3.696044955	33.5
L4					28.4
L5	20170.42370	20170.46279	DEM	-1.937936485	52.5
L6	18772.61160	18773.16783	DEM	-29.62960146	40.1
L7	12206.06290	12206.84246	DEM, ACE	-63.86686734	64.5
	12220.08980	12220.34692	DEM, MET, ACE	-21.04051641	
L9					52.3
L10	17580.43760	17579.99728	DEM	25.046248	31.5
L11	14870.47030	14870.62168	DEM, 9-MET	-10.18017567	31.7
L12	12206.06290	12206.84246	DEM, ACE	-63.86686734	64.5
L13	16918.57380	16918.49658		4.564037188	34.5
L14	13541.06560	13540.92976		10.03207606	28.5
L15	14980.44430	14980.31488		8.639329876	34.7
L16		15365.83597	2 ACE		32.4
L17	14364.62170	14365.25629		-44.17721631	26.8
L18	12769.64490	12770.35352		-55.49285086	19.7
L19	13002.05480	13002.20087	DEM	-11.23453194	36.5
L20	13365.77070	13365.53097	DEM	17.93618979	22.9
L21	11565.05541	11564.49163		48.74857753	44.7
L22					41.8
L23	11199.13960	11199.11761		1.963632992	27
L24	11185.02740	11185.03471	DEM	-0.653641671	52.9
L25	10693.46300	10693.34952		10.61199725	52.1
L26					
L27	8993.28970	8993.271975	DEM	1.970913936	16.5
L28	8875.31060	8875.421703	DEM	-12.51820979	12.8
L29	7273.46450	7273.491698		-3.739345947	46
L30	6410.61260	6410.667045	DEM	-8.492948084	33.9
L31	7871.10070				21.4
L32	6315.19890	6314.651869	DEM	86.62134141	26.3
L33	6254.42280	6254.446543	DEM, MET	-3.796193631	27.3
L34					
L35					
L36					

DEM = N-terminal methionine truncation
ACE = acetylation
MET = Methylation
BMT = beta- methylthiolation and a K to T point mutation

Figure 4.1 presents an example of data from the top-down approach. Shown in Figure 4.1 is the first 15 minutes of the total ion chromatogram of the SmRC strain where all elution of the purified ribosomal sample from the reverse phase separation occurred, along with a deconvoluted mass spectrum corresponding to the chromatographic peak at 2.38 minutes with the isotopic pattern for the component at nominal mass 11184 Da shown. The measured isotopic packet of this species is consistent with the calculated isotopic packet (MAIM) of intact ribosomal protein L24. The measured isotopically resolved peak at 11,184.383 Da was within 8 part per million of the calculated MAIM value of 11,184.286 Da for this protein. For comparison, searching the entire *E. coli* proteome database for measured mass 11,184.383 Da (L24) reveals only one protein within 9.0 Da, and when searching the entire database including a maximum of 5 PTMs and a 2 Da mass window five proteins are found. These proteins, even with modifications, are well within the separable mass range for the FT-ICR-MS.

The bottom-up MS measurements often provide a more extensive list of proteins than the top-down, and indicate the presence of some other components that are consistent across the ribosomal purification process for each of the three strains. The most abundant of these proteins, observed in all three strains, include bacterioferrin observed at 39-63% sequence coverage and 6-8 unique peptides, GroEL with 23-42% sequence coverage and 9-16 unique peptides, and a Co-A linked acetaldehyde dehydrogenase with 35-39% sequence coverage and 25-27 unique peptides.

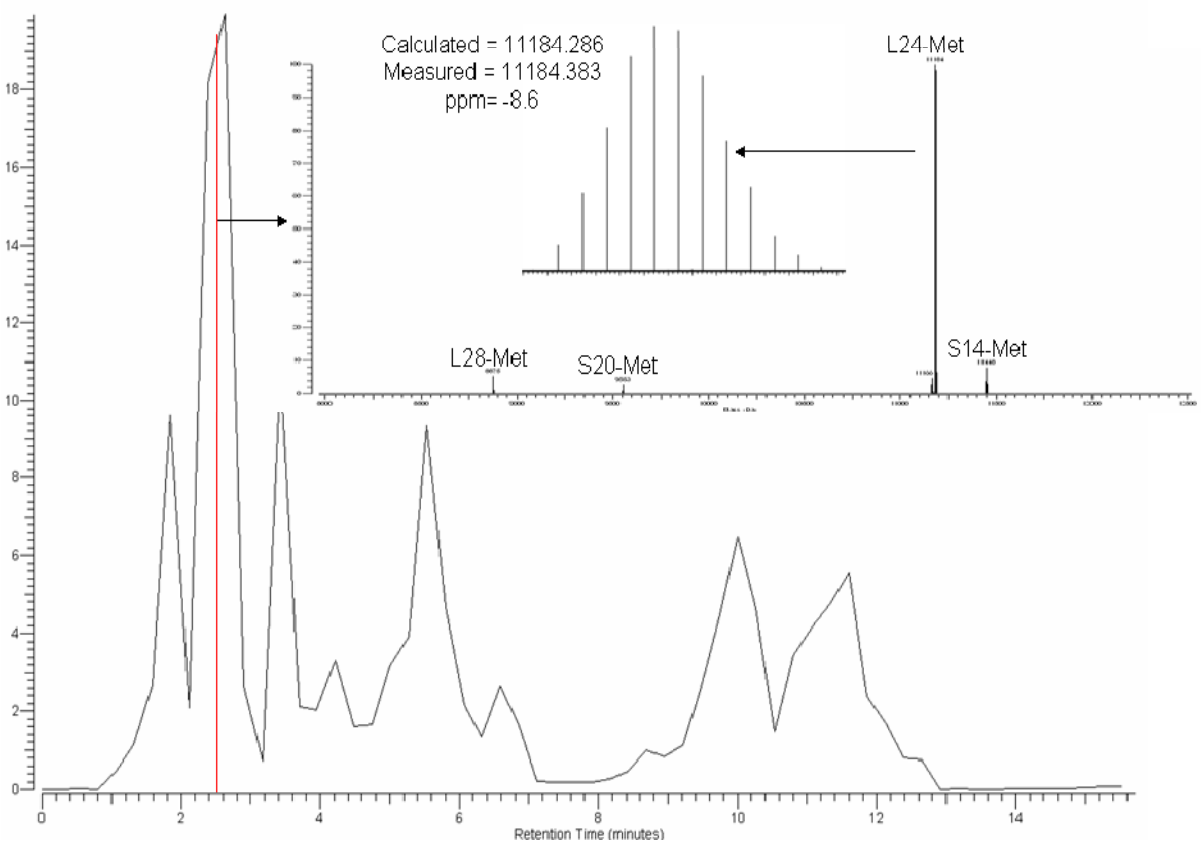


Figure 4.1: 15 minutes of the total ion chromatogram for the SmRC strain. Showing where all elution of the purified ribosomal sample from the reverse phase separation occurred, along with a deconvoluted mass spectrum corresponding to the chromatographic peak at 2.38 minutes with the isotopic pattern for the component at nominal mass 11184 Da shown. The measured isotopically resolved peak at 11,184.383 Da was within 8 part per million of the calculated isotopically averaged value of 11,184.286 Da for the L24 protein. Methionine truncation within the figure is labeled as MET.

Post Translational Modifications of Antibiotic Resistant E. coli Ribosomal Proteins

The integrated top-down and bottom-up approach allows for the identification of PTMs, their location, as well as isoforms of ribosomal proteins. Included in the top-down PTM searches were N-terminal modifications of methionine truncation, methylation, acetylation, and β-methylthiolation. In addition, the bottom-up analysis contained β-methylthiolation of aspartic acid, single acetylations, and mono-, di-, and trimethylated lysines and arginines. All of these modification types have been previously identified in ribosomal proteins from *E. coli* [90, 89, 91, 92, 93]. Phosphorylation is common in eukaryotic ribosomal proteins, although, this has yet to be definitively identified in prokaryotic ribosomal proteins, and was therefore excluded from the subset of modifications searched for [94].

N-terminal Methionine Truncations

N-terminal methionine truncation was the most prevalent PTM identified by top-down analysis. Of the 57 ribosomal proteins, 31 ribosomal proteins from all three strains (WT, SmR, SmRC) were identified to have an N-terminal methionine truncation by top-down analysis in this study [Table 4.2-4.4]. The top-down approach identified an N-terminal methionine truncation if the measured intact mass for a ribosomal protein matched that obtained by subtracting the mass of a methionine residue (131.0405 Da) from the mass calculated from the DNA-derived amino acid sequence. The results of this searching agreed perfectly with previous identified *E. coli* ribosomal proteins with N-terminal methionine truncations [90].

β-methylthiolation

The novel β-methylthiolation PTM is known to occur at the D88 residue of the S12 *E. coli* ribosomal protein [93]. In the top-down analysis of the WT strain, a MAIM molecular mass of 13651.527 was observed corresponding to the S12 ribosomal protein with a β-methylthiolation and an N-terminal methionine truncation with a calculated MAIM mass of 13651.469. These top-down measured and calculated MAIM values for S12 are within a -4.2 ppm mass accuracy. The S12 protein was identified in the bottom-up analysis, although the peptide containing the D88 β-methylthiolation was not observed. Even though the bottom-up search did not yield any positive peptide matches the mass accuracies provided by the top-down measurement still provides strong evidence. The SmR and SmRC strains also contain this modification along with a point mutation and will be discussed later.

Acetylation

A number of ribosomal proteins from the three streptomycin resistant strains of *E. coli* were identified by top-down and bottom-up methods to have an acetylation, including L7, S5, L15, and L16. The L7 protein in *E. coli* is known to have an N-terminal methionine truncation and acetylation of the serine at first position [90, 89]. This modification state was found for L7 in all three strains (WT, SmR, SmRC) of antibiotic resistant *E.coli* in the top-down analysis. The measured MAIM value for the modified L7 protein was 12205.520 and the calculated MAIM value was 12205.502 providing a mass accuracy of -1.5 ppm. The bottom-up analysis did not find the N-terminal peptide for this protein therefore missing the acetylation at the serine.

Both bottom-up and top-down measurements provide conformation of the N-terminal truncation and acetylation of the S5 protein in all three strains. The bottom-up analysis shows the N-terminal peptide acetylated at the alanine for all three WT, SmR, and SmRC strains. Further more, the top-down analysis confirms this with a high mass accuracy of 12 ppm.

Methylation

The S21 protein only in the SmRC strain was found in top-down analysis with 2 isoforms present (Figure 4.2). The first isoform present is S21 with a N-terminal methionine truncation, the second observed isoform within the same mass spectrum is the S21 protein with a N-terminal methionine truncation plus 7 methylations. For the S21 isoform with 7 methylations, the measured MAIM value of 8465.699 was obtained, and when compared to the calculated MAIM value of 8465.806 Da, a ppm of 16 is obtained. However, only two peptides, with low protein sequence coverage, were obtained for S21 in the bottom-up analysis and these two peptides did not contain a methylation. This could be due to peptides with Methylations being lost from this small protein when the trypsin digestion was performed.

Also identified by top-down and bottom-up analysis were the L11, S11, L3, and L33 ribosomal proteins with methylations. The only observed isoform of the L11 protein was found in the SmRC and WT strains with an N-terminal methionine truncation and 9 methylations.

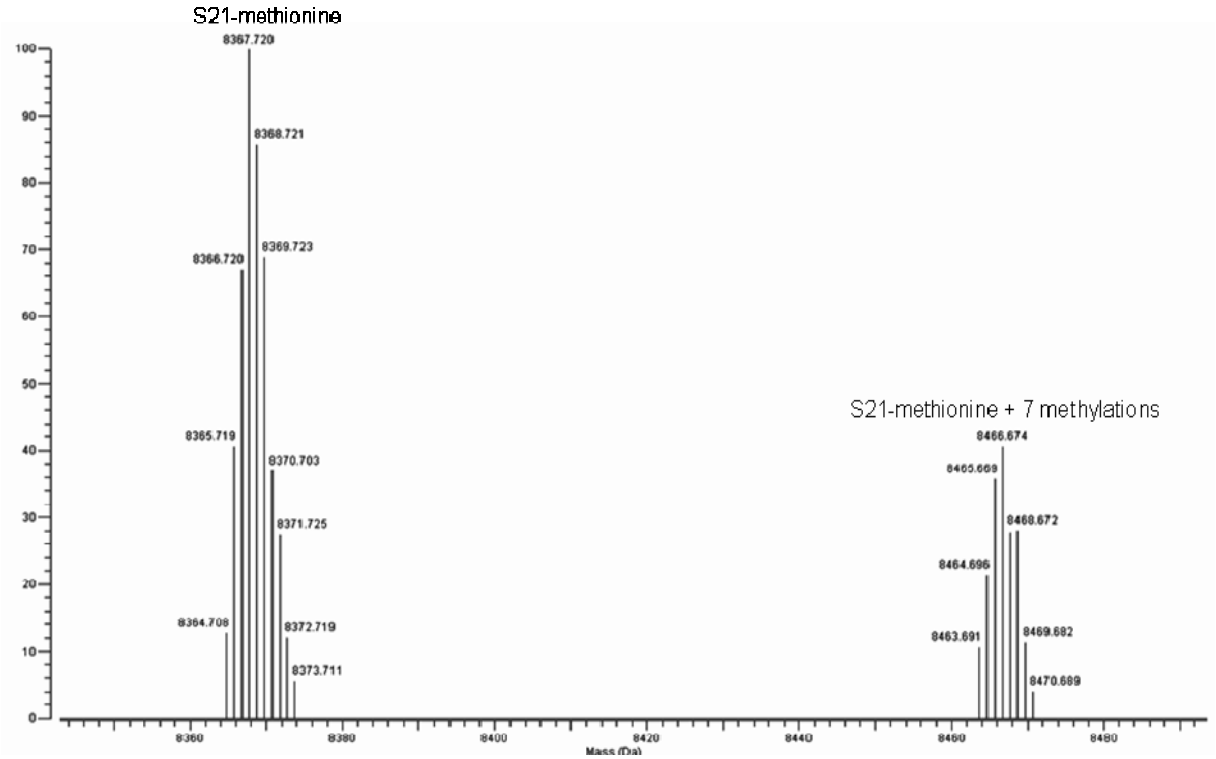


Figure 4.2: The S21 protein in the SmRC strain was found with top-down analysis to have 2 isoforms present. The first isoform present is S21 with a N-terminal methionine truncation, the second observed isoform within the same mass spectra is the S21 protein with a N-terminal methionine truncation plus 7 methylations.

The S11, L3 and L33 proteins were all identified with a single methylation. This data is consistent with previous studies of *E. coli* ribosomal proteins [90].

Point Mutations

The S12 protein was identified only in the SmR and SmRC strain with a N-terminal methionine truncation, β-methythiolation, as well as a lysine 42 to a threonine point mutation (Figure 4.3). The top-down analysis shows the S12 protein with a retention time of 1.58 minutes and a measured MAIM value of 13624.45 providing a ppm of -1.9 when compared to the calculated MAIM value of 13624.42 (Figure 4.3). Searching of the bottom-up data found the S12 protein with the point mutation. Peptide 36-VYTTTPTKPNSALR-49 was found with the threonine in position 42 instead of the lysine (Figure 4.3). The y-ion series is labeled in Figure 4.3 for peptide 36-VYTTTPTKPNSALR-49 with the y8 ion highlighted corresponding to the threonine.

Discussion of Antibiotic Resistant E. coli Results

Two strains of streptomycin resistant *E. coli* (SmR and SmRC) ribosomal proteins were analyzed and compared to the wild type K12 *E. coli* strain in order to see any differential post translational modifications or amino acid substitutions present that may confer streptomycin resistance in *E. coli*. The wild type strain (WT) was used as a baseline to ensure growth, purification, and analysis was consistent for the SmR and SmRC strains. Also the WT strain was used to provide a baseline modification state that the SmR and SmRC strains could be compared to and see how far antibiotic strains vary from the parent strain. In the bottom-up analysis of the WT strain, five ribosomal proteins were not observed including L34, L35, L36, L26 and S22.

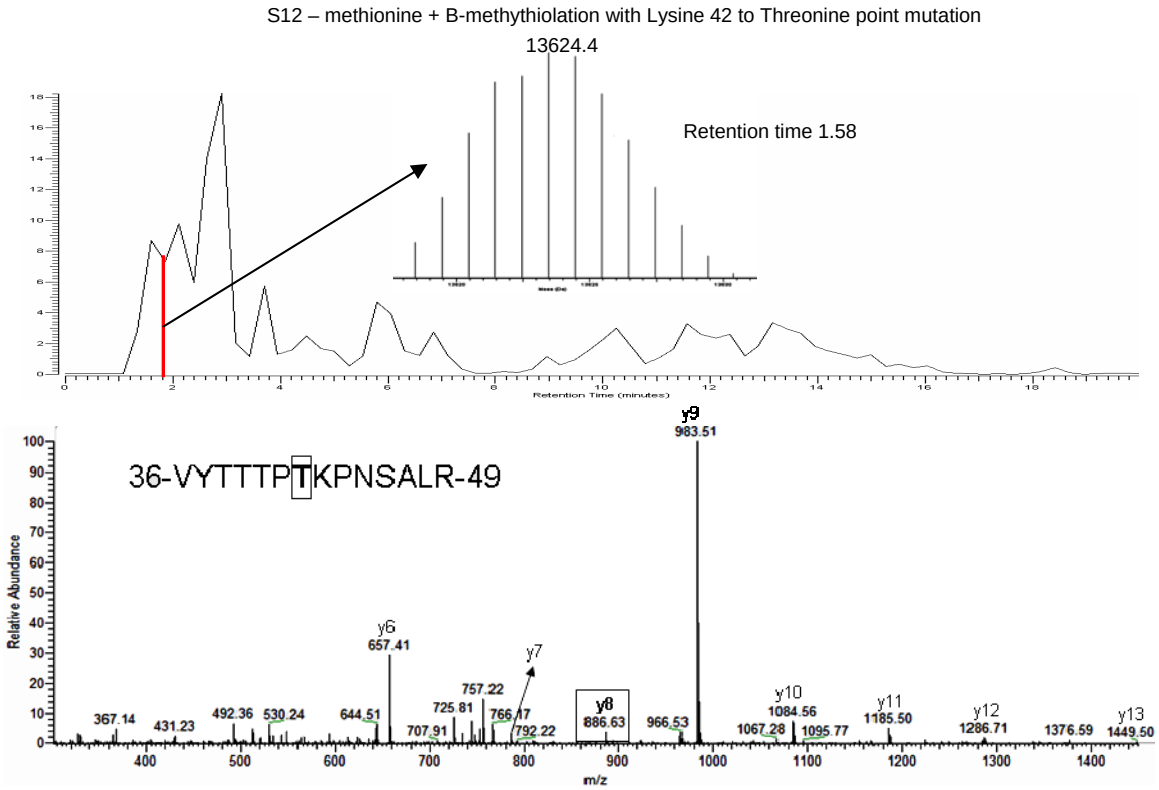


Figure 4.3: Total ion chromatrogram ans MS/MS spectrum for S12. The total ion chromatogram is shown for the SmRC sample with the S12 protein, at retention time 1.58 min, expanded out containing a N-terminal methionine truncation, β-methythiolation, as well as a lysine 42 to a threonine point mutation. The y-ion series is labeled for peptide 36-VYTTTPTKPNSALR-49 with the y8 ion highlighted corresponding to the threonine point mutation.

The L34, L36 and S22 proteins have a high percentage of basic residues providing a large number of trypsin cleavage sites. The L34 protein has 16 lysines and arginines with its 46 amino acid sequence, followed by L36 with 12 and S22 with 11. Due to these sequences being so rich in trypsin cleavage sites, many of the resulting peptides fall below the lower m/z limit for isolation and fragmentation within the mass spectrometer. In the top-down analysis, three of the larger proteins S1, S2, and L4 were not observed. The FT-ICR-MS has sufficient mass range to analyze these proteins, but prior experience with intact protein chromatography indicates that larger species such as these three proteins are difficult to elute off an on-line C4 reverse phase column, under the top-down experimental conditions employed. Low abundance and different hydrophobicities may prevent proteins from being observed due to irreversible binding to the reverse phase column, and lower than detectable concentrations.

When the total number of proteins observed in both the SmR and SmRC strains are compared, a higher number of ribosomal proteins are observed for the SmRC strain. A total of 41 ribosomal proteins were observed in the top-down analysis for the SmR strain as compared to 43 for the SmRC strain. Also, five more ribosomal proteins were able to be identified for the SmRC strain in the bottom-up analysis. These observed differences in the two strains could be due to the compensation that was allowed to occur for the SmRC strain. The acquired resistance to streptomycin by *E.coli* has an associated fitness cost resulting in slowed growth. The compensated derivative strain (SmRC) was obtained by evolving an isolate of the original streptomycin resistant strain (SmR) through repeated serial passage, in the laboratory, until it had "compensated" for the reduced fitness and recovered a wild type comparable growth rate. This "compensation"

process in the SmRC strain may allow for higher protein expression providing, a greater number of identifiable proteins. This compensation within the SmRC strain is thought to come from intra- or extra-genic mutations and differential post translational modifications that stabilize the resistance phenotype in the population. However, a process of differentially post translationally modifying ribosomal proteins to compensate fitness is thought to occur, what these modifications are have been difficult to obtain by traditional molecular techniques. Using top-down mass spectrometry, differentially expressed modifications could be examined for the SmR and SmRC strains. An example of these differential post translational modifications is the S21 and L16 ribosomal proteins. The S21 ribosomal protein is present in the SmRC strain with 7 methylations and a N-terminal methionine truncation (Figure 4.2), whereas in the SmR strain S21 only contains the N-terminal methionine truncation and no identified methylations. The L16 ribosomal protein was also identified with a differential post translation modification within the SmRC strain and not in the SmR strain. The L16 protein was identified with two acetylations in the SmRC strain, which were not identified in the SmR strain. The use of top-down mass spectrometry provided, for the first time, a way of examining differential post translational modifications in "compensated" streptomycin resistant strains of *E. coli*.

Streptomycin resistance within *E. coli* is thought to occur from point mutations within the ribosomal proteins. One such previously identified point mutation is the lysine 42 to threonine in ribosomal protein S12. The S12 protein is known as the "hinge" protein in the ribosomal complex, and plays an important role in the structural

conformation [95]. Therefore, it is understandable why this protein would be an important target for antibiotic resistance.

The integrated top-down and bottom-up method identified the S12 protein with a N-terminal methionine truncation, β-methythioaltion, as well as a lysine 42 to a threonine point mutation in the SmR and SmRC strains. The identification of the S12 ribosomal protein with the lysine to threonine point mutation only within the SmR and SmRC strains, and not the WT strain, provides further conformation toward its role in streptomycin resistance.

Conclusions

Employing the integrated top-down and bottom-up approach, first allowed for a comprehensive evaluation of ribosomal proteins from *R. palustris* and second of antibiotic resistant strains of *E. coli*. The analysis of component proteins of the 70S ribosome from *R. palustris* enhanced several aspects of the analysis. The intact protein measurements include the aggregate contribution of all modifications to the protein, allowing for the discrimination of isoforms with different molecular masses, while the peptide data provided the location of the modification in many instances.

Not only was this method useful in the analysis of *R. palustris* ribosomes; the use of integrated top-down and bottom-up mass spectrometry approaches provided insight into the role of ribosomal proteins in streptomycin resistance in *E. coli.* The identification of differential modifications may provide starting points for future biological analysis of antibiotic resistance within bacterial species.

Chapter 5

Evaluation of PTMs and Isoforms in Protein Complexes from *Rhodopseudomonas palustris* for Key Regulation Sites

All of the data presented below is accepted for publication Heather M. Connelly, Dale A. Pelletier, Tse-Yuan Lu, Patricia K. Lankford, and Robert L. Hettich. Characterization of pII Family (GlnK1, GlnK2, GlnB) Protein Uridylylation in Response to Nitrogen Availability for *Rhodopseudomonas palustris. Analytical Biochemistry, Accepted, In Press* (2006). *All MS sample preparation, experiments and data analysis were performed by Heather M. Connelly.*

Introduction

The analysis of protein complexes, and their associated PTMs, that play a key role in regulation was an important aspect in the development of this dissertation work. This analysis allowed for the improved identification of PTMs and protein complexes that was need for future work in the analysis of multiple growth states from *R. palustris* (Chapter 7).

The movement of ammonium across biological membranes is a process that is conserved throughout all domains of life from bacteria to man [96]. In bacteria, the pII family generally plays a pivotal role in nitrogen metabolism regulation due to its ability to sense internal cellular ammonium concentrations [96,97,98]. This protein family is able to sense and transduce an ammonium signal, via protein-protein interactions, to a variety of enzymes involved in nitrogen metabolism [99,100,101]. The pII proteins GlnK and GlnB in *Escherichia coli* are trimers that functions as small signal transduction proteins and are able to sense the status of cellular nitrogen within prokaryotic cells [102]. The crystal structure of GlnK in *E. coli* has a compact barrel structure around 50Å in diameter and 30Å high, with an unstructured T-loop protruding from the upper surface

[96,103]. Residue tyrosine-51 at the apex of the T-loop is uridylylated in nitrogen starved cells, with the process being reversed when nitrogen is sufficient [96].

The general nitrogen regulation system (ntr), which has been most extensively studied in *E. coli*, controls the transcriptional activity of a number of genes involved in nitrogen regulation and assimilation, such as *glnA* (encoding glutamine synthetase) and *nifA* (encoding the transcriptional activator for the other nif genes) [104,105]. In *E. coli*, there are two levels of regulation involving the uridylylation of both the GlnB and GlnK proteins. The first level within the cascade is the uridylylation (under low ammonium conditions) and de-uridylylation (under high ammonium conditions) of the GlnK protein in direct response to the intracellular nitrogen concentration, which in turn regulates the AmtB ammonium transporter's movement of ammonium across the cell membrane [96,106,107]. This regulation of AmtB occurs when the cellular nitrogen status reaches a certain level, and de-uridylylation of GlnK allows for the direct binding of AmtB and sequestration to inhibit further ammonium transport [96,108,109]. The second level in the cascade is the uridylylation of the GlnB pII functional protein, which is thought to play a role in regulating enzymatic activity of glutamine synthetase (GlnA), which catalyzes the conversion of glutamate to glutamine, by controlling the level of adenylylation on tyrosine-397. Adenylytransferase (AT) is the enzyme that adenylylates and deadenylylates GlnA. Regulation as to which reaction the adenylytransferase catalyzes is determined by either unmodified GlnB (which stimulates adenylylation of glutamine synthetase) or uridylylated GlnB (which stimulates deadenylylation of glutamine synthetase) [110,111,112] (Figure 5.1).

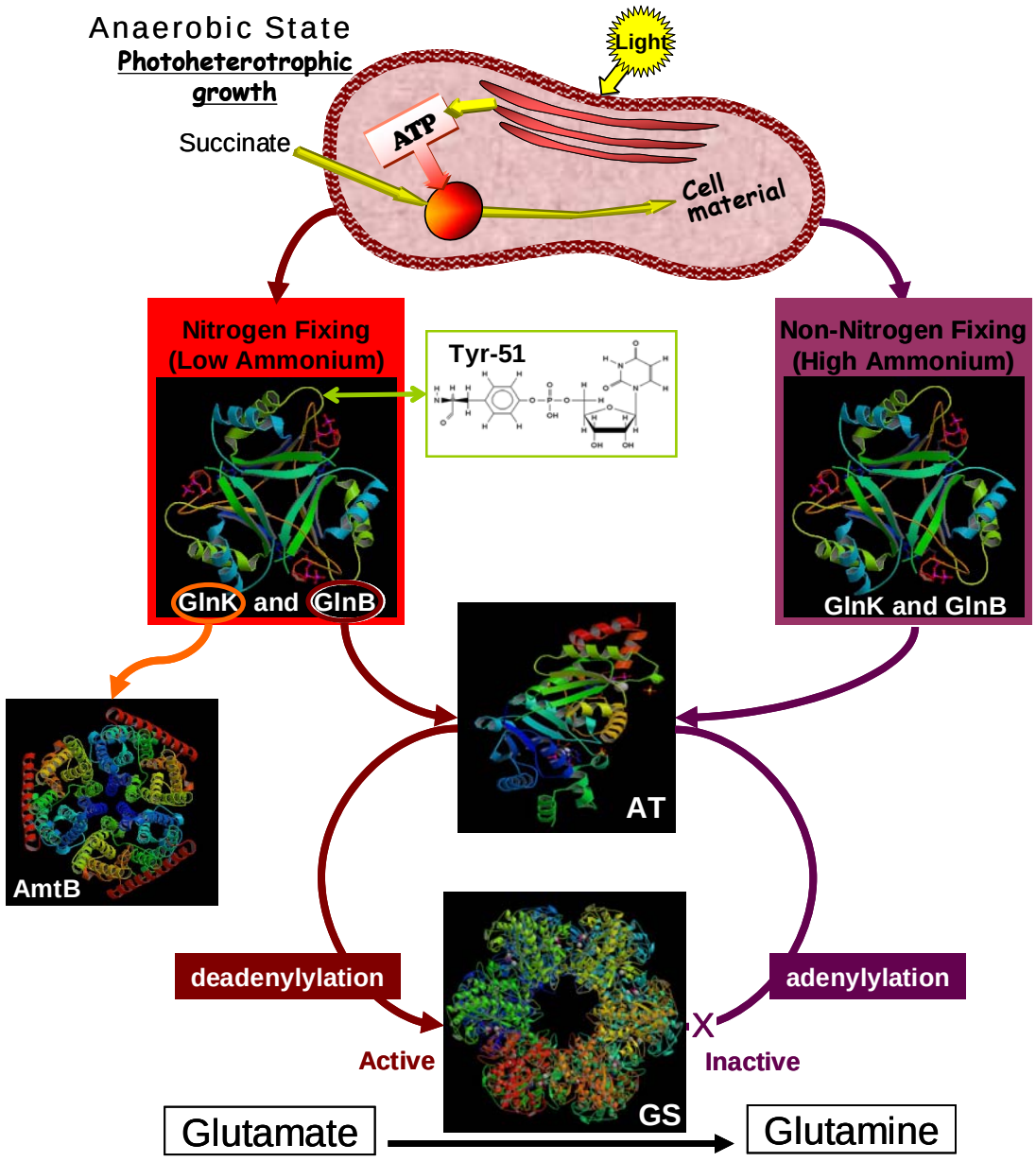
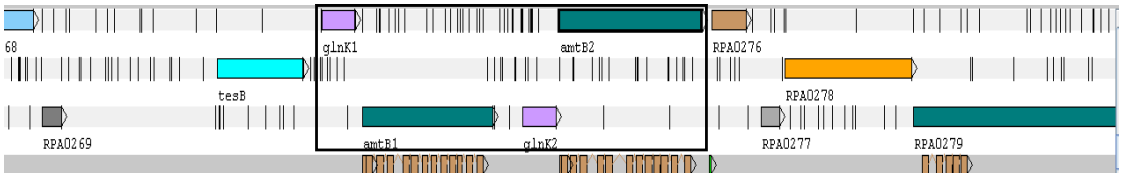


Figure 5.1: Proposed model for glutamine synthetase (GS) regulation in *R. palustris* based on known models in *E. coli*. Two metabolic states were interrogated in this study. The growth state shown on the left is anaerobic, grown in the light without oxygen (photoheterotrophic) with no ammonium present (nitrogen fixing conditions). The growth state shown on the right is photoheterotrophic growth with ammonium sulfate present in the growth media (non-nitrogen fixing). The tagged GlnK and GlnB proteins are suspected to be uridylylated on Tyr-51 under nitrogen-fixing conditions which in turn activates the adenylytransferase (AT) to deadenylylate glutamine synthetase (GS), Under non-nitrogen fixing growth the lack of uridylylation on Tyr-51 leads to the inactive form of GS . Figure adapted from Larimer et al. *Nature Biotechnology*, 2004.

In the purple non-sulfur anoxygenic phototrophic bacterium *Rhodopseudomonas palustris,* the GlnK proteins also are expected to function as a primary regulator point in ammonium sensing and thus regulation of the glutamine synthetase pathway. However, *R. palustris* has unique metabolic versatility in its modes of energy generation and carbon metabolism, and unlike *E. coli*, it is able to thrive under severe nitrogen limiting conditions by fixing atmospheric nitrogen. As such, it is possible that *R. palustris* may utilize a nitrogen-ammonium regulation system that varies from other commonly studied bacteria such as *E. coli* [13, 11]. In *R. palustris* there are three encoded forms of pII proteins; GlnK1 (*RPA0272*), GlnK2 (*RPA0274*), and GlnB (*RPA2966*). Also unique in *R. palustris* is the encoding of two AmtB transporters within the same operon as GlnK1 and GlnK2, with each transporter corresponding to one of the two GlnK proteins (Figure 5.2). Based on information from other bacteria, it is likely that uridylylation of these proteins is a key aspect of controlling the glutamine synthetase pathway. Figure 5.1 outlines a proposed pathway in which GlnK2 is the uridylylation target protein that is sensitive to the availability of nitrogen, and the uridylylation of GlnB activates glutamine synthetase.

Traditionally, nitrogen-ammonium regulation has been evaluated with immunoblotting and native gel analysis for a number of organisms, such as *Escherichia coli*, *Gluconacetobacter diazotrophicus*, *Rhodospirillum rubrum*, and *Rhodobacter capsulatus* [96, 97, 110, 111,113]. In this methodology, a series of plasmids are created to evaluate the effects of removing or altering a gene, or series of genes, to look at protein expression and modification levels. Once the protein or proteins are isolated, there are two common approaches employed, including Western blotting using antibodies or native gels to look at the migration differences of modified or unmodified proteins of interest.

(A)



(B)

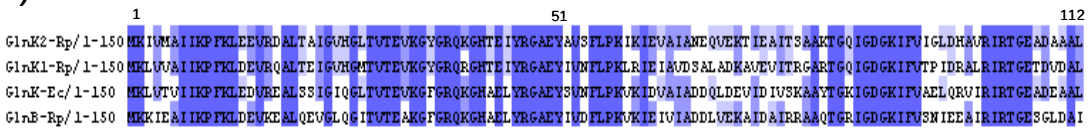


Figure 5.2: Artemis view and sequence alignment for GlnK1, GlnK2, and GlnB. (2A) The Artemis view showing *R.palustris* genome, with the bold box highlighting the GlnK1, GlnK2 as well as both AmtB transporters located within the same operon. (2B)The sequence alignment and homology comparison of the *R. palustris* (labeled Rp) GlnB, GlnK1, GlnK2 proteins, along with the *E. coli* (labeled Ec) GlnK protein as a comparison are shown.

While these methodologies are informative, they are also very labor intensive and sometimes difficult to reproducibly perform. Therefore, using these methods in tandem with mass spectrometry is able to provide a comprehensive technique for the examination of protein modifications quickly and accurately in most cases.

Mass spectrometry is a rapidly emerging tool for protein identification and characterization. Intact protein or top-down mass spectrometry can be used to characterize the GlnK and GlnB proteins, as well as their modification state, to ascertain the level of regulation in the glutamine synthetase pathway of *R. palustris*. This powerful method can provide information on the natural state of intact proteins, including details about post-translational modifications (PTM's), truncations, mutations, signal peptides, and isoforms, due to top-down mass spectrometry's ability to measure the molecular weight of a protein very accurately and detect any covalent modifications that alter the mass of a protein [114]. This information is often difficult to obtain by the more common peptide or "bottom-up" mass spectrometry methods, where intact proteins are digested with a protease such as trypsin or Glu-C and the resulting peptide mixtures are analyzed by MS or MS/MS methods. The top-down mass spectrometry approach was first introduced with electrospray-Fourier transform ion cyclotron resonance mass spectrometry (ESI-FTICR-MS) [22, 23, 24]. The dynamic range, sensitivity, and mass accuracy offered by high performance FTICR-MS afford not only unambiguous protein identification in many cases, but also detailed information about protein modifications.

In this report, we will focus on the investigation of the GlnK and GlnB proteins modification state for *R. palustris* as a function of nitrogen availability to the growing bacterial cultures, and ultimately glutamine synthetase activation or inactivation. This

will be achieved by isolating affinity-tagged GlnK2, GlnK1, and GlnB complexes under nitrogen-fixing and non-nitrogen fixing growth conditions. Using this method in conjunction with an integrated high resolution top-down and bottom-up mass spectrometry approach should reveal detailed information about the presence and isoforms of GlnK and GlnB proteins. In particular, uridylylation of GlnK is suspected to be a key regulatory aspect of nitrogen availability, while the uridylylation of GlnB is thought to play a key role in the regulatory aspect of glutamine synthetase. Both of these modifications states of GlnK and GlnB should be identifiable in the different growth samples. The experimental section can be found in Chapter 2.

Results

Characterization of GlnK1, GlnK2, and GlnB Under Non- Nitrogen Fixing Conditions

GlnK2

Affinity purifications of the GlnK2 protein complex from *R. palustris* under non-nitrogen fixing growth conditions were performed in order to examine the baseline modification state of the complex and associated proteins. A Western blot was obtained for the GlnK2 protein complex after affinity purification using antibodies to the 6X his-tag present within the GlnK2 protein. This immunoblot shows only the GlnK2 protein band at approximately 13 kDa present on the gel (Figure 5.3). Lane one of the immunoblot is the whole cell extract, followed by lane three which contains protein extract obtained after the nickel purification step and finally lane five contains the final V5 antibody purification extract from the affinity purification (Figure 5.3).

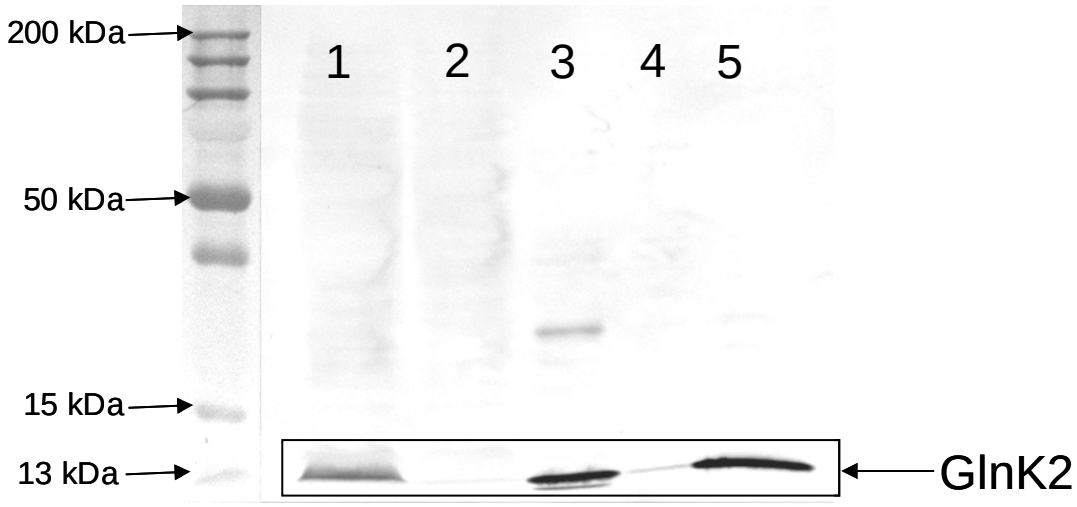


Figure 5.3: Western Blot of GlnK2 complex at approximately 13 kDa. Lane one of the immunoblot is the whole cell extract, followed by lane three which contains protein extract obtained after the nickel purification step and finally lane five contains the final V5 antibody purification extract from the affinity purification. Lanes two and four contain the bypass wash elute.

Lanes two and four are bypass wash steps that are tested as a control to ensure no protein loss is occurring in the wash steps. The Western analysis provides another way of visualizing and confirming the protein tagging procedures and affinity purifications for the GlnK2 protein.

Under these growth conditions, the affinity purification procedure for tagged GlnK2 yielded abundant proteins corresponding to the unmodified GlnK1 and unmodified, tagged GlnK2, as determined by top-down ESI-FTICR-MS measurements (Figure 4A). The m/z values for both the unmodified GlnK1 and tagged GlnK2 are present within the FTICR mass spectra, with GlnK1 having a higher intensity than the tagged GlnK2. GlnK1 is seen with a charge state package ranging from +12 to +15 with the two most abundant m/z values of 883.9309 and 951.8519, while the tagged GlnK2 charge state package ranges from +17 to +20 with the two most abundant m/z values being 859.9071 and 907.6828, providing a distinguishing charge state series for both proteins. The measured and theoretical molecular mass values of unmodified GlnK1 (*measured 12,360.824 Da, calculated 12,360.776 Da, 4 ppm mass error*) and unmodified tagged GlnK2 (*measured 16,318.980 Da, calculated 16,318.859 Da, 7 ppm mass error*) agree very well, and demonstrate the power of this high-resolution mass spectrometric technique. To demonstrate the ability and considerations of high-resolution top-down mass spectrometry for directly identifying *R. palustris* proteins, the entire *R. palustris* proteome database was queried with the measured molecular mass of 12,360.824 Da. The closest match, as illustrated above, was the unmodified GlnK1 protein; the only other protein within a window of 5 Da was the RPA4690 hypothetical protein at 12, 360.828 Da. Since the focus of this work is to examine modified proteins, this search then was

expanded to include a range of common post-translational modifications. In particular, this measured molecular mass of 12,360.824 Da was searched against the entire *R. palustris* proteome database, this time including methionine truncation, as well as any possible combination of 0-5 methylations, acetylations, oxidations, and disulfide bonds. This search yielded only six possible proteins within a 2 Da window. Obviously, the availability of peptide or MS/MS data for intact proteins would greatly help limit the search space.

The experimental determination of only unmodified GlnK1 and GlnK2 is consistent with the expectation that under non-nitrogen fixing growth conditions, the high ammonium levels within the cell leads to an inactive form of AmtB, and thus there is no need to modify the GlnK2, since it is only expressed endogenously at a very low level [102]. The forced over-expression of this tagged GlnK2 under non-nitrogen fixing conditions explains why this affinity purification yields only the tagged version of this protein.

The same sample used to generate Figure 5.4 was examined by the bottom-up MS technique, in which proteolytic digestion was used to generate peptides for LC-MS/MS interrogation. The bottom-up experimental results confirmed the top-down data. Under these non-nitrogen fixing conditions, the peptide MS results verified the presence of unmodified GlnK1 at 89.3% sequence coverage with 16 unique peptides and unmodified, tagged GlnK2 at 94.6% sequence coverage with 25 unique peptides. The bottom-up MS measurements often are more extensive than the top-down, and indicate the presence of some other minor components in this sample.

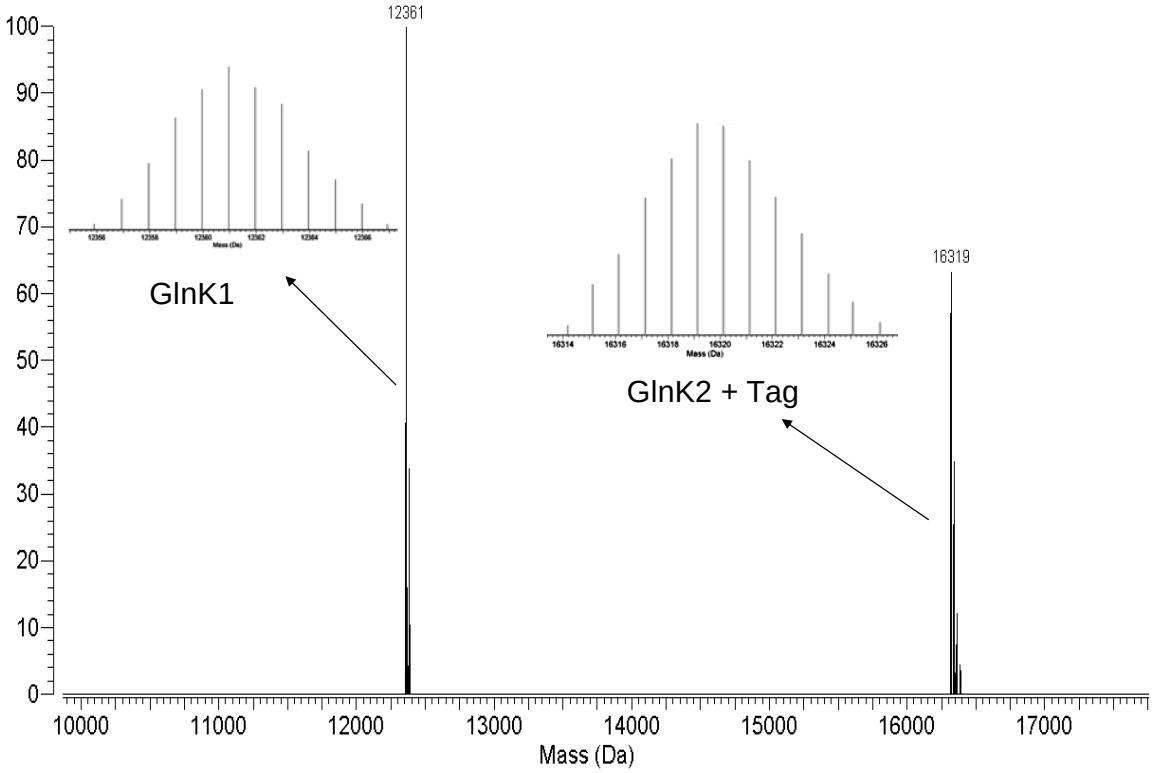


Figure 5.4: ESI-FTICR mass spectrum of GlnK2 affinity purification from *R. palustris* grown under non-nitrogen fixing conditions. Only unmodified GlnK1 and unmodified tagged (tag refers to 6X His-tag and V5 antibody tag) GlnK2 proteins are present in the growth state.

A possible DNA-binding protein Hu-alpha (RPA 2953) and the GlnB regulatory protein were each observed with 2-peptide hits and sequence coverage's of about 30%. The low abundance of the GlnB protein not observed in the LC-MS/MS experiments could be due to the weak affinity of GlnB associated with GlnK2 in the affinity purification. A few other species were detected with single peptide hits, but were not considered to be significant enough for confident identification.

GlnK1

Also performed, were affinity purifications of the tagged GlnK1 protein complex from *R. palustris* under non-nitrogen fixing growth conditions, in order to examine the baseline modification state of the complex and associated proteins. As determined by top-down ESI-FTICR measurements for the GlnK1 affinity purifications, two forms of GlnK1 were identified [Figure 5.5]. These two isoforms of GlnK1 correspond to the tagged and untagged forms of GlnK1. Both of these proteins were identified with a 5-10 ppm mass accuracy. It is important to remember that the 6X His tag and V5 antibody tag used for affinity purifications are inserted within the plasmid DNA. Therefore, the untagged version of the protein is coming from the bacterial chromosomal DNA. The experimental determination of only unmodified GlnK1 in both the tagged and un-tagged versions is consistent with the expectation that under non-nitrogen fixing growth conditions, the high ammonium levels within the cell leads to an inactive form of the AmtB transporter, and thus there is no need to modify the GlnK1, since there is no need for the cell to transport ammonium.

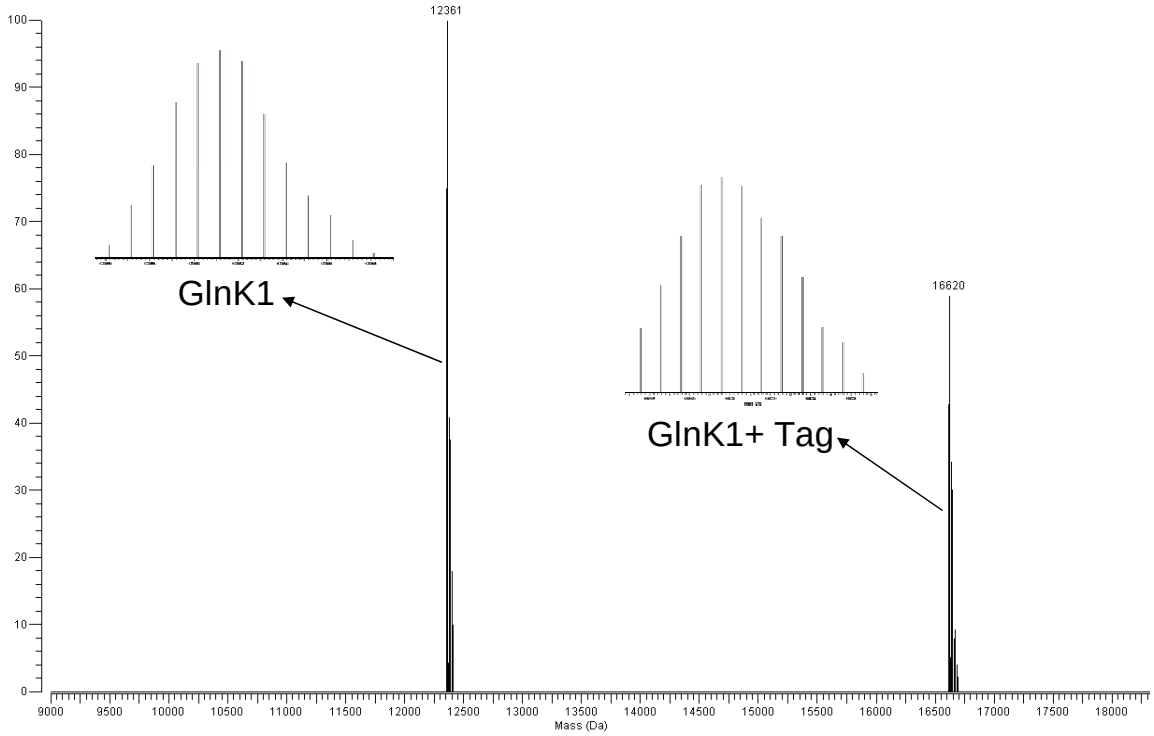


Figure 5.5: ESI-FTICR mass spectrum of GlnK1 affinity purification from *R. palustris* grown under non-nitrogen fixing conditions. The tagged and un-tagged versions of the protein are present.

The expression of GlnK1 under non-nitrogen fixing growth conditions will yield both tagged and untagged versions of this protein due to GlnK1 existing as a multimeric form, which interacts with both GlnK2 and GlnB [102]. Within the non-nitrogen fixing growth state, GlnK2 is expected to be expressed at a significantly lower level in the cell as compared to GlnK1. This observation is supported by bottom-up data, where GlnK1 is present at 35.7% and 7 unique peptides, while GlnK2 has only one unique peptide and 21.4% sequence coverage.

GlnB

Affinity purifications of the GlnB protein complex from *R. palustris* under non-nitrogen fixing growth conditions were also performed. Using top-down ESI-FTICR-MS to examine the GlnB affinity purifications reveals two isoforms of GlnB present within the non-nitrogen fixing growth state [Figure 5.6]. The two isoforms of GlnB identified correspond to the tagged and untagged versions of GlnB. The experimental determination of only unmodified GlnB is consistent with the expectation that under non-nitrogen fixing growth conditions, the high ammonium levels within the cell leads to an inactive form of glutamine synthetase, and thus there is no need to modify the GlnB. Again it is not surprising to see the un-tagged version of GlnB since it is coming from the chromosomal DNA of *R. palustris*. Bottom-up data indicates that GlnB is present at 30% sequence coverage and 4 unique peptides.

The top-down data shown in Figures 5.4-5.6, along with the bottom-up information, indicate that the expression of the tagged, GlnK1, GlnK2, and GlnB proteins and subsequent affinity purification procedures are effective in enriching the targeted sample for mass spectrometric characterization.

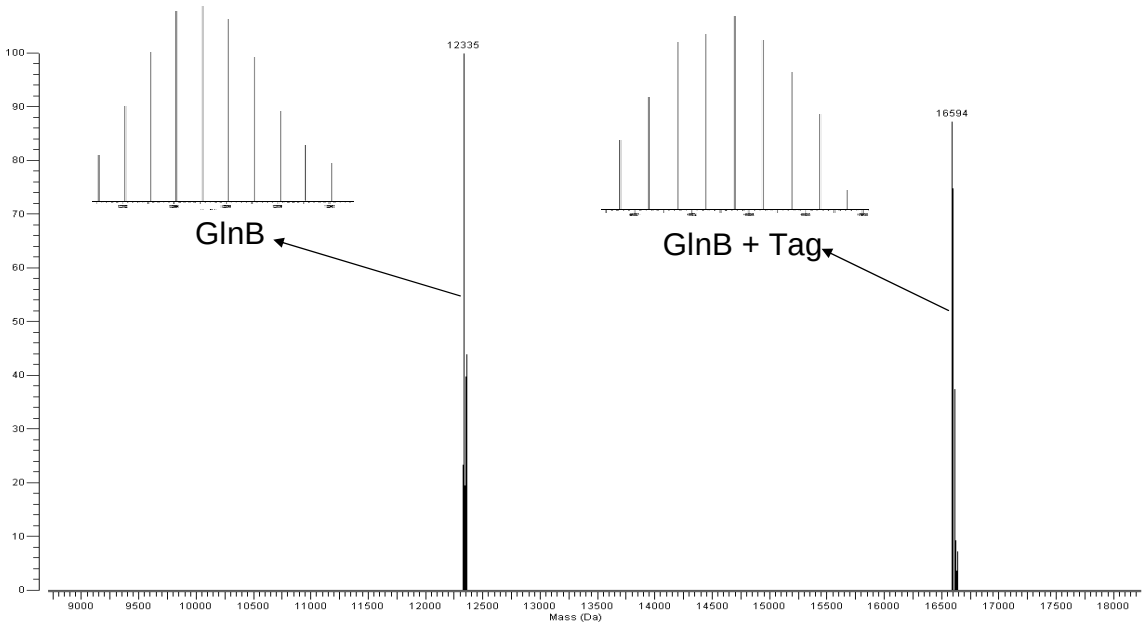


Figure 5.6: ESI-FTICR mass spectrum of GlnB affinity purification from *R. palustris* grown under non-nitrogen fixing conditions. The tagged and un-tagged versions of the protein are present.

Thus, this approach can be used for *in-vivo* studies of GlnK1, GlnK2, and GlnB modifications as a function of growth state. The observation of GlnK1 in the affinity purification of the tagged GlnK2 verifies the robustness of this method. The absence of significant quantities of other proteins (i.e. non-specific binding) also attests to the potential of this affinity method.

Characterization of GlnK1, GlnK2, and GlnB Under Nitrogen Fixing Conditions

GlnK2

Affinity purifications of GlnK2 from *R. palustris* grown under nitrogen fixing conditions revealed the presence of four isoforms of expressed GlnK2, as shown in Figure 5.7. As expected, the unmodified tagged version of GlnK2 is present; however, it also is accompanied by the uridylylated version of the tagged protein, as well as the untagged GlnK2 and the uridylylated untagged version of this protein. While the presence of untagged GlnK2 initially may be surprising, it is important to remember that GlnK2 is thought to exist in a trimeric form, which interacts with both GlnK1 and GlnB [97]. Thus, the expression of GlnK2 under these growth conditions will yield both tagged and untagged versions of this protein. The affinity purification targets the tagged GlnK2, which will bring-down the other components of the protein complex. Once again, the high mass accuracy of 3-5 ppm afforded by the ESI-FTICR-MS provides the ability to confirm the molecular masses of all four isoforms by comparing them with the calculated masses.

In the direct infusion ESI-FTICR-MS experiments of the GlnK2 affinity purification shown in Figure 5.7, GlnK1 was not observed, even though it was detected under non-nitrogen fixing conditions (Figure 5.4).

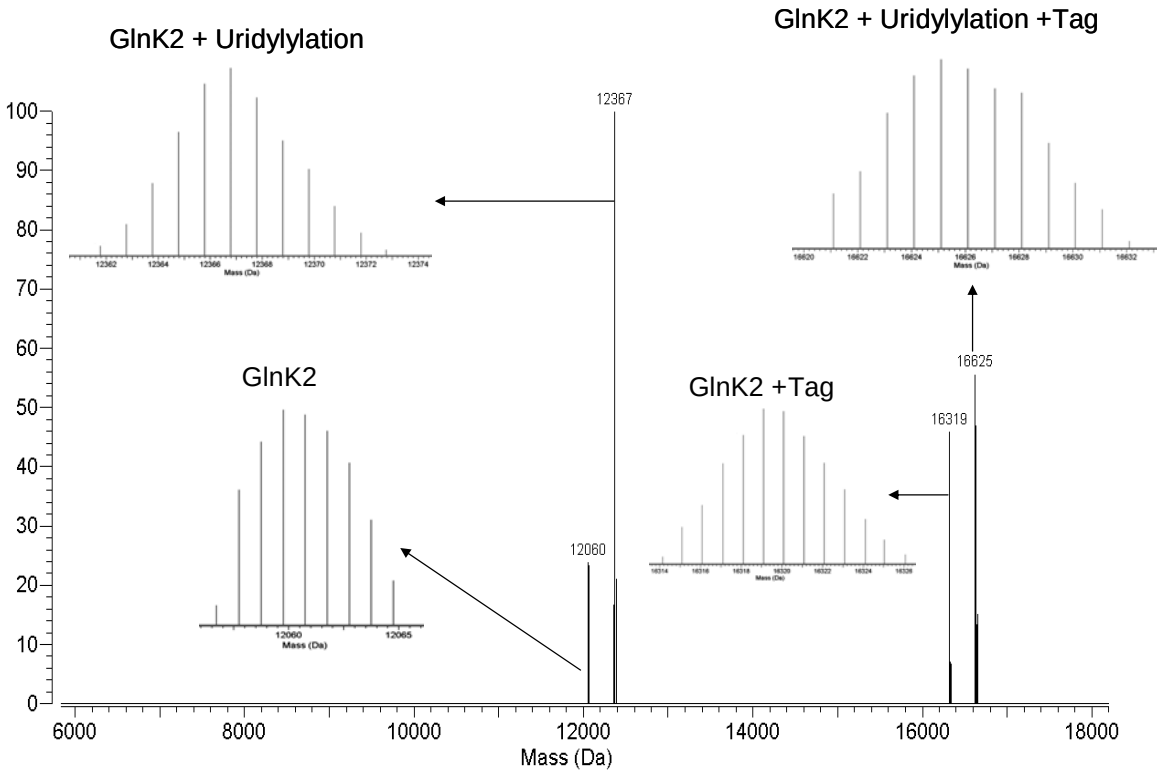


Figure 5.7: ESI-FTICR mass spectrum of GlnK2 affinity purification from *R. palustris* grown under nitrogen fixing conditions. Four different isoforms of the GlnK2 protein are present within the growth state, including the tagged and untagged isoforms of the protein as well as the modified and unmodified isoforms.

This absence of GlnK1 in Figure 5.7 most likely is due to the overwhelming amount of uridylylated untagged GlnK2 at similar molecular mass. This ionization suppression effect in direct infusion ESI-MS experiments is not uncommon, and is one of the major reasons for performing LC-MS measurements (i.e. to provide spatial separation of proteins prior to measurement).

In order to evaluate whether ionization suppression was a factor in GlnK1 detection, samples of GlnK2 affinity purifications from the nitrogen fixing and non-nitrogen fixing growth conditions (confirmed to contain GlnK1) were mixed at a 1:1 ratio. Even though GlnK1 was observed in Figure 5.4, in this set of experiments from the mixed sample, GlnK1 was not observed, therefore supporting the ionization suppression postulation. To alleviate this problem, an online liquid chromatography ESI-FTICR-MS experiment was performed to search for GlnK1 in the associated affinity purification complex of the nitrogen fixing sample. This LC-FTICR-MS experiment allowed for partial chromatographic separation of the GlnK2 (all isoforms) from GlnK1, and provided evidence that GlnK1 was present in this sample (Figure 5.8). Even though the chromatographic separation of GlnK1 was incomplete from GlnK2, there was distinct evidence for unmodified GlnK1, as well as all four isoforms of GlnK2 in this sample. The GlnB protein was not observed in the LC-FTICR-MS or ESI-FTICR-MS experiments; this could be due to the low abundance or weak affinity of GlnB associated with GlnK2 in the affinity purification. The first peak eluting in the chromatogram is ubiquitin, added as an internal standard.

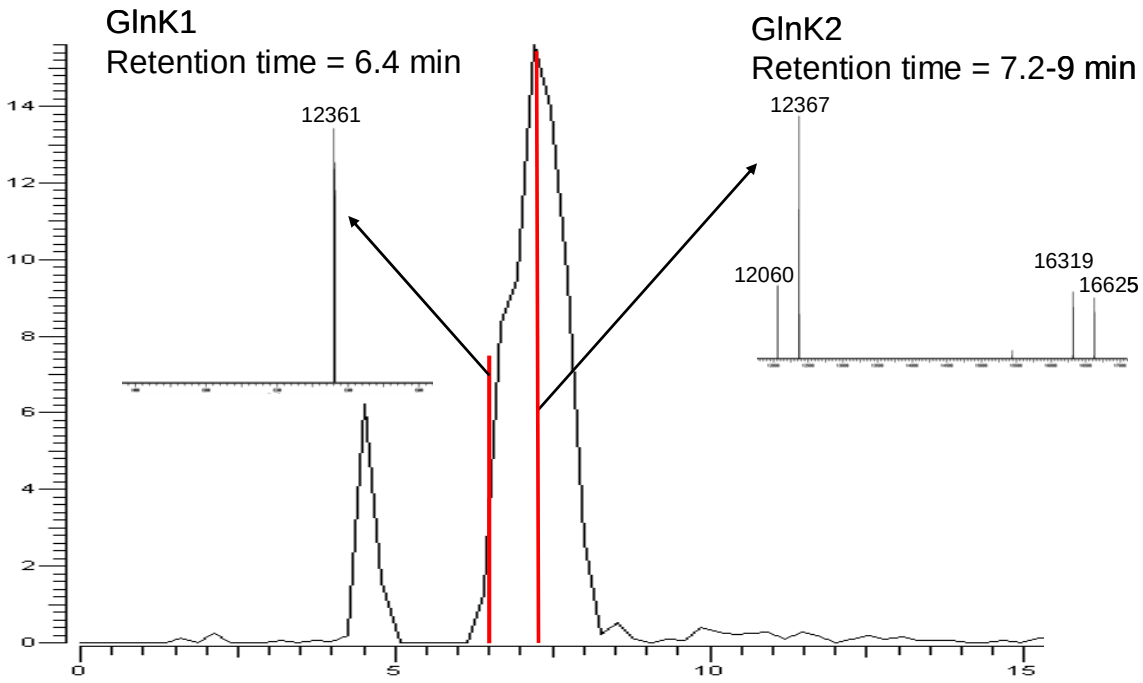


Figure 5.8: LC-FTICR-MS total ion chromatogram of GlnK2 affinity isolation showing the GlnK1 protein as well as all four forms of the GlnK2 protein. The first peak in the chromatogram (t_r = 4.5min) is ubiquitin which was used as an internal standard for the chromatography.

Bottom-up MS characterization of this nitrogen-fixing sample confirmed the overwhelming presence of GlnK2 (100% sequence coverage with 20 unique peptides) and GlnK1 (92.9% sequence coverage with 15 unique peptides). As before, the bottom-up MS measurements indicated the presence of some other minor components in this sample. The same DNA-binding protein Hu-alpha (RPA 2953) and the GlnB regulatory protein were each detected at a fairly low level. A few other species, primarily ribosomal proteins, were detected with single peptide hits, but were not considered to be definitive enough for identification. Bottom-up analysis also confirmed the presence of two unique uridylylated peptides from the GlnK2 complex under nitrogen fixing growth conditions. Each of these two peptides contains tyrosine- 51, which was suspected to be the uridylylation site in GlnK2. Peptides 48-GAEY*AVSFLPK-58 and 48-GAEY*AVSFLPKIK-60 were present with high DBDigger scores, of 58.9 for a +2 and 34.0 for a +1, and abundant ion intensities. The MS/MS spectrum of peptide 48-GAEYAVSFLPK-58 is shown in Figure 5.9, with the b and y fragmentation ion series labeled within the mass spectrum. This MS/MS spectrum shows unambiguously that the tyrosine residue within the peptide contains the uridylylation, which adds a mass shift of 306.02 Da. Inspection of all the other peptides failed to reveal a tyrosine uridylylation at any other position. Also present in the MS/MS were GlnK2 peptides containing tyrosine-51 that were *not* modified with the uridylylation. These finding are consistent with the observation of the unmodified form of GlnK2 in the top-down mass spectra. There were no uridylylated peptides found for GlnK1 within the GlnK2 affinity purification, verifying the top-down data that this protein is unmodified.

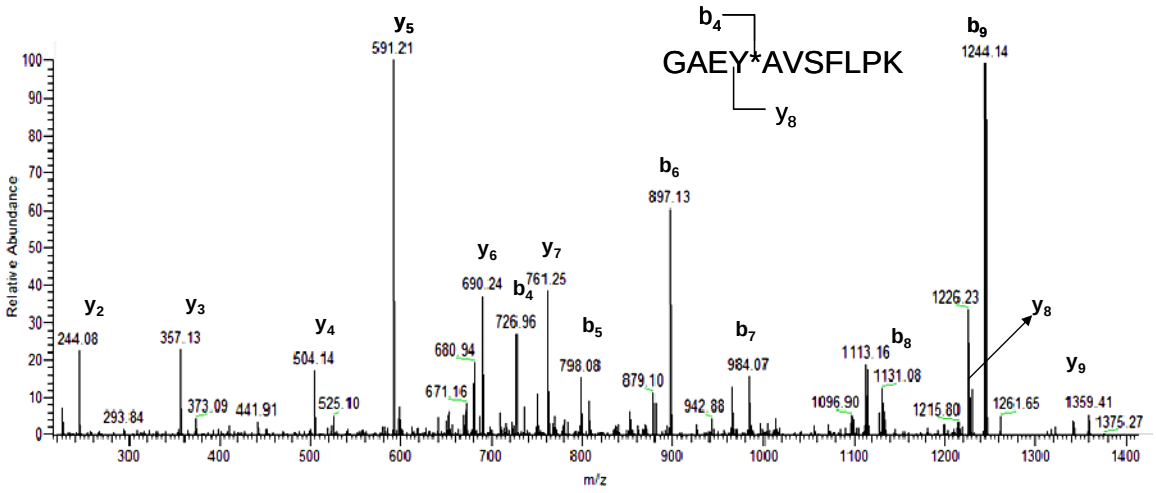


Figure 5.9: MS/MS spectrum of uridylylated peptide 48-GAEY*AVSFLPK-58. The spectrum has the b and y ions labeled showing the uridylylation on tyrosine 51 (y_8 and b_4 ions).

Even though the bottom-up MS measurements verified the presence of both uridylylated Y-51 and non-uridylylated Y-51 peptides for GlnK2, this technique did not provide further details about which of the four isoforms was present or whether all were present. Top-down analysis of the GlnK2 modification state proved to be the most valuable and efficient way of confirming the presence of the multiple isoforms.

By combining the top-down and bottom-up mass spectrometry approaches, it was possible to determine not only the unique site of uridylylation in GlnK2, but also the range of isoforms present under nitrogen-fixing growth conditions.

In both the LC-FTICR-MS and ESI-FTICR-MS experiments of the GlnK2 affinity purification, GlnK2 under nitrogen fixing growth conditions was observed to be uridylylated in this affinity purification. Therefore, GlnK2 seems to play a key role in the regulation of nitrogen availability in *R. palustris* and activation of the AmtB ammonium transporter. This mechanism of AmtB regulation is different from other well characterized systems such as *E. coli* where the primary regulation site is GlnK1. Also *R. palustris* differs from other bacterial species in that it encodes three pII proteins providing additional regulation sites for the bacteria within the glutamine synthetase pathway.

GlnK1

Affinity purifications of GlnK1 from *R. palustris,* grown under nitrogen fixing conditions, revealed the presence of two isoforms for both GlnK1 and GlnK2, as shown in Figure 5.10. As expected, the unmodified tagged version of GlnK1 is present; however, it also is accompanied by the uridylylated version of the tagged protein, as well as the unmodified GlnK2 and the uridylylated untagged version of this protein.

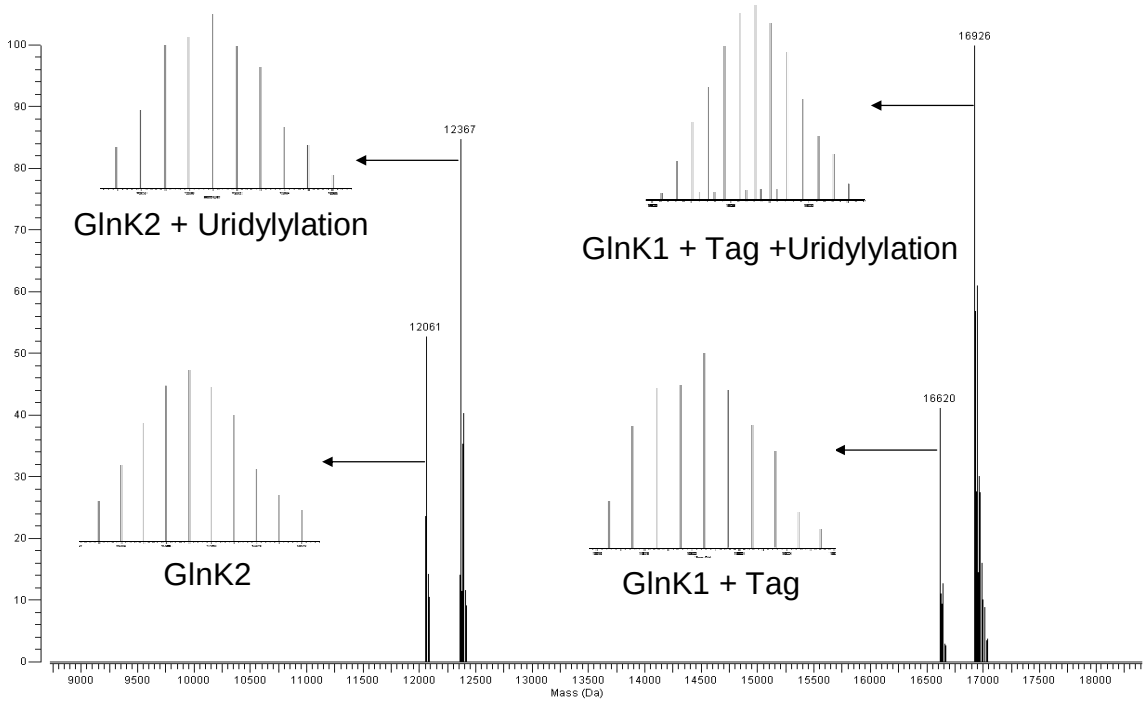


Figure 5.10: ESI-FTICR mass spectrum of GlnK1 affinity purification from *R. palustris* grown under nitrogen fixing conditions. The unmodified and modified tagged (tag refers to 6X His-tag and V5 antibody tag) GlnK1 isoforms are present in the growth state as well as the unmodified and modified isoforms of GlnK2.

The presence of untagged and modified GlnK2 again is due to the multimeric form of the complex [102]. Thus, the expression of GlnK1 under these growth conditions will yield both tagged and untagged versions of this protein. The affinity purification targets the tagged GlnK1, which will bring down the other components of the multimer. This mixture of GlnK1 and GlnK2 isoforms from this affinity purification may indicate that under nitrogen fixing conditions GlnK2 plays a more primary role in the regulation of AmtB. This is also supported by the lower abundance of GlnK1 in the GlnK2 affinity purification under nitrogen fixing conditions. Once again, the high mass accuracy afforded by the ESI-FTICR-MS provides the ability to confirm the molecular masses of all the GlnK1 and GlnK2 isoforms by comparing them with the calculated masses. Bottom-up data further confirms the top-down data with GlnK1 present at 73.2% sequence coverage and 14 unique peptides. The GlnK2 protein is present at 92.0% sequence coverage and 19 unique peptides within the same affinity purification. Bottom-up analysis also confirmed the presence of two unique uridylylated peptides (48-GAEY*IVNFLPK-58 and 41-GHTEIYRGAEY*IVNFLPK-58) for GlnK1 as well as the unique uridylylated peptide 48-GAEY*AVSFLPK-58 from GlnK2.

GlnB

Affinity purifications of GlnB from *R. palustris,* grown under nitrogen fixing conditions, revealed the presence of four isoforms of expressed GlnB, as shown in Figure 5.11. This affinity purification yields the unmodified tagged version of GlnB which is also accompanied by the uridylylated version of the tagged protein, as well as the untagged GlnB and the uridylylated untagged version of this protein.

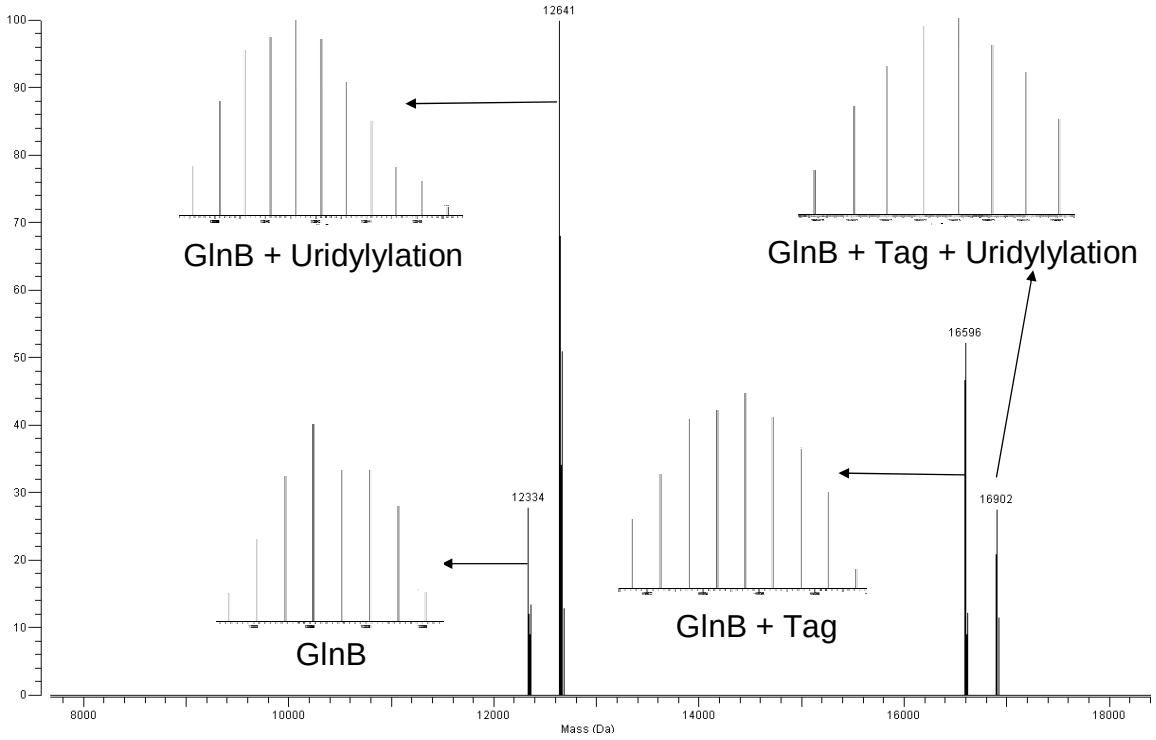


Figure 5.11: ESI-FTICR mass spectrum of GlnB affinity purification from *R. palustris* grown under nitrogen fixing conditions. Four different isoforms of GlnB proteins are present in the growth state, including the tagged and untagged isoforms of the protein as well as the modified and unmodified isoforms.

It is again important to remember that GlnB exist in a trimeric form, therefore yielding all four isoforms of the GlnB protein. Bottom-up MS data identifies GlnB at 50.7% sequence coverage and 11 unique peptides. In the ESI-FTICR-MS experiments of the GlnB affinity purification, only GlnB under nitrogen fixing growth conditions was observed to be uridylylated. Therefore, GlnB may be a key regulation site of glutamine synthetase in *R. palustris*. This mechanism of glutamine synthetase regulation is similar to other well characterized systems such as *E. coli* where the primary regulation site is GlnB.

Wild Type *R. palustris* GlnK1, GlnK2, and GlnB Analysis

In order to examine if the plasmid constructs that forced over-expression of GlnK1, GlnK2, and GlnB would alter the normal state of the proteins, wild type *R. palustris* cells were analyzed for the presence and modification state of the GlnK1, GlnK2, and GlnB proteins. Wild type cells were grown under both photoheterotrophic conditions as well as nitrogen fixing conditions, the cells were lysed and total protein was extracted, FPLC anion exchange fractionation performed, and the resulting fractions were analyzed by LC-FTICR-MS and LC-MS/MS. GlnK1and GlnB proteins were detected in the cells grown under photoheterotrophic conditions, both in the unmodified forms. GlnK2 was not observed under these conditions. For this growth condition, the most abundant protein was GlnB, which may be indicative of endogenous expression under non-nitrogen fixing conditions. The absence of GlnK2 expression and modification is consistent with the growth state of the cells, which is anaerobic non-nitrogen fixing. The bottom-up experimental results for the non-nitrogen fixing growth conditions revealed GlnK1 at 38.4% sequence coverage and 3 unique peptides, and GlnB at 51.8% sequence

coverage and 7 unique peptides. No peptides were observed for GlnK2, in accord with the top-down MS results.

Under nitrogen fixing growth conditions, the GlnK2 and GlnB proteins were identified in both the unmodified and modified states. The GlnK1 protein was not observed under nitrogen fixing conditions. This observation is consistent with the observations in the affinity purifications of higher GlnK2 expression under nitrogen fixing conditions. These results provide evidence that plasmid constructs did not alter the natural state of the complex.

Conclusions

The pII proteins, GlnK1, GlnK2, and GlnB, all appear to play an essential role in ammonium and nitrogen regulation for *R. palustris*. Affinity purifications, in conjunction with top-down mass spectrometry, permitted the isolation and characterization of the functional state and isoforms for these proteins as a function of nitrogen availability. Under non-nitrogen fixing conditions, all of these pII proteins are unmodified. Under endogenous growth conditions, GlnB and GlnK1 are abundant, whereas GlnK2 was not observed. Under nitrogen fixing conditions, all of these pII proteins are uridylylated, all on the Tyr-51 positions. Thus, pII protein uridylylation appears to be tightly coordinated with nitrogen availability. The presence of tagged and untagged protein isoforms also provided evidence for the multimeric conformations of these species, thereby supporting results obtained from *E. co*li that these proteins exist in trimers.

From this work, we conclude the GlnK2 is predominantly expressed and uridylylated in *R. palustris* under nitrogen limited conditions, presumably to regulate the AmtB transporter. Unmodified GlnK1 is abundant in non-nitrogen fixing conditions, but

also is uridylylated under nitrogen-fixing conditions. As expected, GlnB is expressed as an unmodified protein under non-nitrogen fixing conditions, while it also is uridylylated under nitrogen-fixing conditions, likely regulating glutamine synthetase. By comparing endogenous growth vs. affinity labeling conditions, we determined that the plasmid construction did not alter the normal state of the proteins, suggesting that this experimental protocol can be used to probe the natural modification conditions of such proteins.

In this study, top-down mass spectrometry using FTICR-MS was found to be an invaluable tool for determining the post translational modifications on the pII family proteins, GlnK1, GlnK2, and GlnB, in *Rhodopseudomonas palustris*. By using a combined technique of protein affinity purifications and mass spectrometry it was determined, for the first time, that GlnK2, GlnK1 and GlnB proteins possess an uridylylation under nitrogen fixing growth conditions in *R. palustris*. This information allowed for a previously un-afforded glimpse into the modifications and isoforms of the proteins that regulate the AmtB transporter and glutamine synthetase in *R. palustris*.

Chapter 6

Computational Searching Algorithms Developed for Integrated Top-down and Bottom-up Data for the Identification of PTMs

All of the data presented below is in preparation for submission Heather M. Connelly, Robert L. Hettich, Chandrasegaran Narasimhan, Gary J. VanBerkel, Vilmos Kertesz. Integrated Top-down and Bottom-up Protein and PTM searching: "PTMSearch Plus" *Analytical Chemistry* (2006) *All MS sample preparation, experiments, biological knowledge behind programming, and final data analysis were performed by Heather M. Connelly. Programming was performed by Vilmos Kertesz, post doc in OBMS group.*

Introduction

One of the largest challenges in developing a top-down proteomics platform was the development of a functional proteome informatics capability. At the start of this dissertation, the ProSight [115] and PROCLAME [116] algorithms had been available for the analysis of intact protein and their MS/MS spectra against protein databases as well as PTM prediction. But no major effort had been made to integrate top-down analysis with traditional enzymatic bottom-up analysis for protein identification and PTM analysis.

Integrating "top-down" and "bottom-up" MS-based proteomic strategies provides a powerful tool to examine complex protein mixtures, such as proteins in multi-component complexes or even complete proteomes. An integrated top-down and bottom-up approach allows for a more comprehensive characterization of protein complexes due to the unique strength of each technique. In an integrated approach, intact protein masses from the top-down analysis corresponding to a particular PTM or isoform are able to be compared to the comprehensive list of proteins provided by the bottom-up analysis. This correlation between the two methods can provide PTM location and identity with more

certainty. The comprehensiveness of this technique has been previously demonstrated in studies of the *Shewanella oneidensis* proteome as well as the 70S ribosomal complex from *Rhodopseudomonas palustris* [54, 35].

Current software searching tools can provide very good identifications from top-down protein data, as well as make predictions of possible PTMs on a protein. Examples of these tools are ProSight PTM [115] and PROCLAME [116]. ProSight PTM [115] combines a number of search engines and browser environments into a web application that allows the user to analyze top-down data from proteins in the >10kDa size range. This program uses intact protein masses and fragmentation masses from the intact proteins to provide protein and PTM identifications. This method works well, although, it requires the use of top-down dissociation methods such as infrared multiphoton dissociation (IRMPD) and electron capture dissociation (ECD) that are not available to all labs and may not be as comprehensive for complex mixtures as bottom-up methods employing an enzymatic digestion. The PROCLAME algorithm uses intact protein mass measurements to determine sets of putative protein cleavage and modification events to account for the measured protein masses observed [116]. PROCLAME provide a good prediction algorithm but is unable to incorporate mass spectrometry (MS/MS) data within the process.

Our ORNL developed algorithm PTMSearch Plus is the first software providing a comprehensive search method that allows for the integration of top-down protein identification with the bottom-up peptide data to identify proteins and their associated PTMs [Figure 6.1]. The software is built around multiple instrumentation platforms and data inputs. These multiple instrumentation and data platforms include bottom-up ion

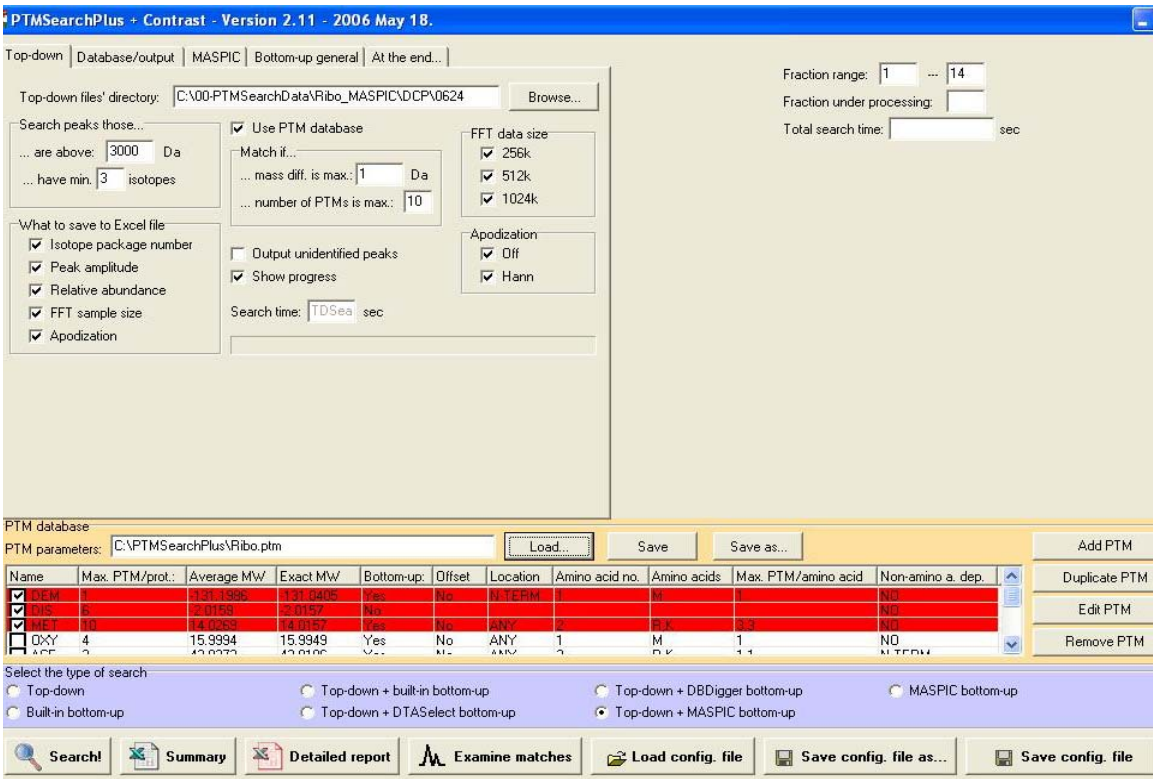


Figure 6.1: Screen shot of *PTMSearch Plus* main data input screen.

trap data, as well as top-down high resolution data such as FT-ICR data. The software can perform independent top-down or bottom-up searches, as well as these two parts of the program being able to interact. By combining these two search capabilities, the results from the top-down search can limit the number of the proteins that are used to generate the database used for the bottom-up search (search time decrease) and in return, the results of the bottom-up search can be used as a confirmation for the proteins with associated PTMs found in the top-down search. This integration reduces the search time dramatically, allowing the user to search for more PTMs on proteins and peptides during a reasonable time frame. The power of this integrated search method is demonstrated using data from analysis of a protein standard mixture and a complex *Rhodopseudomonas palustris* ribosomal protein mixture.

Methods and Software

System Requirements.

PTMSearch Plus was developed using Delphi 3 computer language (Borland Software Corp., Scotts Valley, CA) under Microsoft© Windows XP Home Edition (Microsoft Corp., Redmond, WA) operation system and can be run in any 32-bit Windows environment with at least 256 MB RAM. Currently, the program is free to use for any government or educational institute.

Methodology.

PTMSearch Plus currently supports seven search options allowing the user to perform:

- a standalone "top-down" search
- a standalone internal "bottom-up" search

- MASPIC [117], “bottom-up” search
- an integrated “top-down” and “top-down predicted” internal “bottom-up” search with PTM/peptide limitation
- an integrated “top-down” and “top-down predicted” MASPIC “bottom-up” search with PTM/peptide limitation
- an integrated “top-down” and “top-down predicted” external “bottom-up” search (e.g. accomplished by DBDigger [118] or Sequest, etc.)
- integration of “top-down” search results with already-made DTASelect [61] “bottom-up” search result files

These search options are discussed in details below.

Defining a PTM.

PTMSearch Plus allows for the user to define any number and kind of PTMs without any restrictions. When a PTM is defined, the following parameters must/can be specified: (a) a 3-letter unique ID that is used to identify the PTM; (b) maximum number of the specific PTM that a protein can have; (c) average mass of the PTM (used in top-down search); (d) exact mass of the PTM (used in bottom-up search); (e) if the PTM is used in the bottom-up search (e.g. disulfide bond formation is not used in bottom-up search); (f) "offset": if the specific PTM should be used when calculating the precursor ion's mass but should be removed when calculating the mass of the fragment ions (e.g. it allows the user to search for labile PTMs such as phosphorylation that is removed from tyrosine, histidine, and serine in the ion trap prior/during fragmentation); (g) location of amino acid that is modified (C-, N-terminal or any location on the peptide); (h) number of amino acids that can have the specific PTM; (i) amino acids that can have the specific

PTM; (j) maximum number of the PTM that the amino acids can have (e.g. in the case of methylation; it can be 3 for arginine and lysine, or set to less) and (k) if the C- or N-terminal amino acid can have the PTM independent from the type of the amino acid (e.g. acetylation can occur on the N-terminal amino acid beside every arginine and lysine residues). The user may define and save many PTMs into a dataset and is still able to select a sub-group of PTMs that are used in the actual search.

Standalone "Top-Down" Search

Prior to performing a standalone "top-down" search (Figure 6.2), the user has to generate a peak list for each spectra from the raw experimental data. A Visual Basic script was written that extracts out all the average mass peaks (calculated by the IonSpec software) from across a selected region of the ion chromatogram, and saves them to a file with DCP extension (deconvoluted peaks). From the DCP files that have been generated, the user is able to select the DCP files to be searched through. The user is able to accomplish searching using the same searching conditions on an unlimited number of DCP files in one software run without user intervention by selecting a directory containing the DCP files of interest. Also needed, prior to starting the top-down search, is a specified FASTA protein database; giving the user the ability to search against any annotated organism, combination of organisms or subset of proteins. The user also has the ability to limit the top-down search to the deconvoluted peaks which meet certain criteria: (a) the *m/z* of a deconvoluted peak must be larger than the preset threshold value (3000 Da by default); (b) minimum number of isotopic peaks deriving a deconvoluted peak (3 by default);

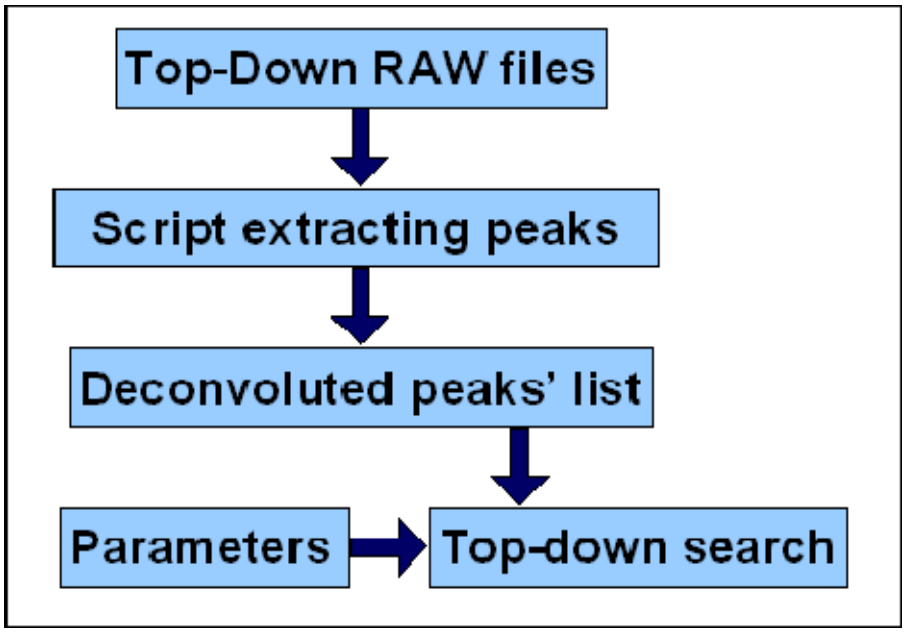


Figure 6.2: Flow chart of the top-down searching method within PTMSearch Plus.

(c) FFT size and (d) apodization settings used to generate the spectra resulting in the deconvoluted peak must be selected in the program. Also, the user has the ability to specify the conditions for a match to the deconvoluted peak found in a top-down spectrum: (a) maximum difference between the *m/z* of a deconvoluted peak and that of the calculated *m/z* of the PTM protein, and (b) maximum number of PTMs on the protein.

Standalone "Bottom-Up" Search

Built into the PTMSearch Plus program is a simple internal bottom-up searching algorithm that is based on the presence and intensity of b- and y-ions in the spectrum. The scoring mechanism is not discussed here in more detail as it was implemented only to demonstrate the power of the integrated top-down and bottom-up searches, such as how limiting the number of peptides to search against in a bottom-up run can drastically reduce search time. PTMSearch plus is designed to allow for the user to use a scoring algorithm of their choice at any time. To demonstrate the ability to implement different scoring algorithms, we implemented the MASPIC scoring algorithm [117] within the software. When using either the internal or the MASPIC standalone bottom-up search, the user must (a) define the number of missed tryptic cleavages; (b) maximum number of PTM a peptide can have (see below); (c) minimum and (d) maximum mass of the tryptic peptide; (e) minimum number of amino acid residue a peptide must have (to exclude short peptides from the search that are normally not unique for a protein); (f) the maximum difference between the peptide and the precursor ion's mass to search the corresponding MS/MS spectra against the b and y fragment ions of the peptide, and (g) the maximum difference between the *m/z* of a peak in the MS/MS spectrum and that of the b and y ions of the peptide to be used in the scoring.

Integrated "Top-Down" and "Bottom-Up" Search

Figure 6.3 shows the simplest approach to integrate "top-down" and "bottom-up" searching algorithms in general. In this case, "top-down" and "bottom-up" data are searched independently and the results are compared. This approach is considered to be a complete search, as all proteins (and their possible PTMs) are checked against the two different datasets. Figure 6.4 shows a different approach, that is implemented in PTMSearch Plus to integrate "top-down" and internal or MASPIC "bottom-up" search algorithms. First, a "top-down" search is accomplished, followed by assigning all combination of possible PTMs found to that particular protein. E.g. if *protein 1* was found with three different PTMs in the "top-down" search: 2 methylations; 4 methylations and a β-methythiolation; then all possible combinations of these PTMs are assigned to *protein 1*. The assigned PTM represents the most complex set of PTMs that a single peptide of the given protein can have. At this point the user has two options: I.) creating peptide sequences exclusively from the proteins found in the "top-down" search using their individually assigned PTMs, or II.) creating peptide sequences from the proteins found in the "top-down" search using their individually assigned PTMs as well as from proteins not found in the "top-down" search using their intact (non-modified) sequence. Each method may drastically decrease the number of peptide sequences used in the "top-down predicted" (i.e. peptide sequences are generated based on the results of the "top-down" search) "bottom-up" search.

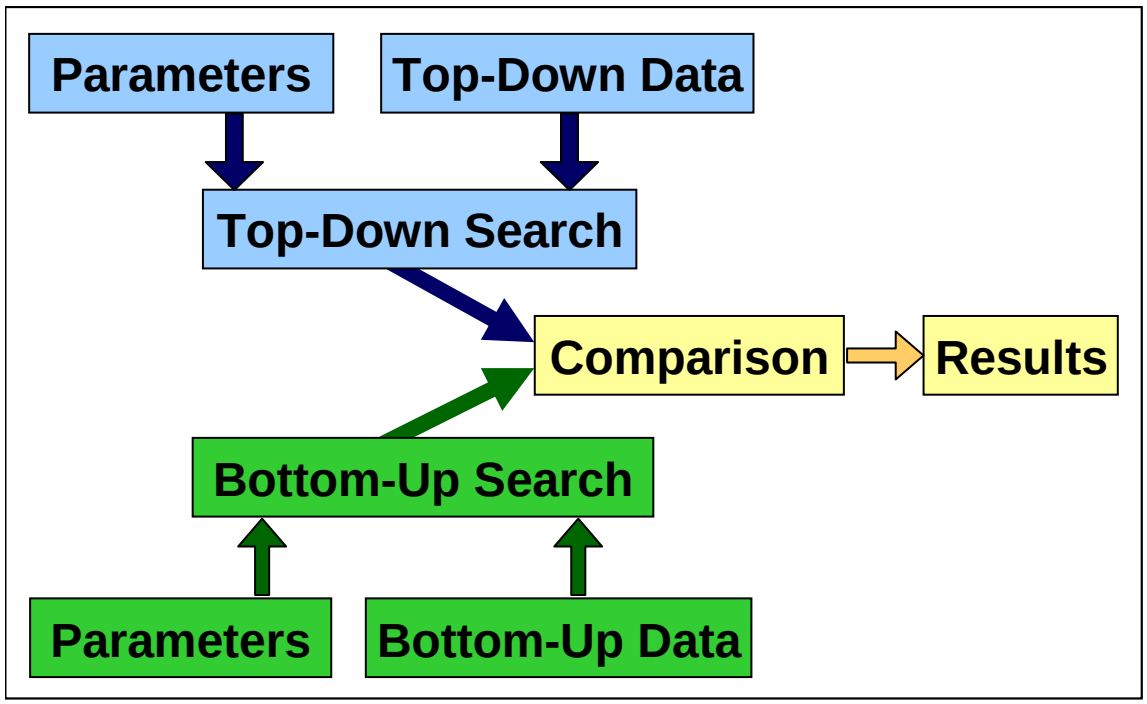


Figure 6.3: Flow chart of simple integration of independent "top-down" and "bottom-up" searching algorithms.

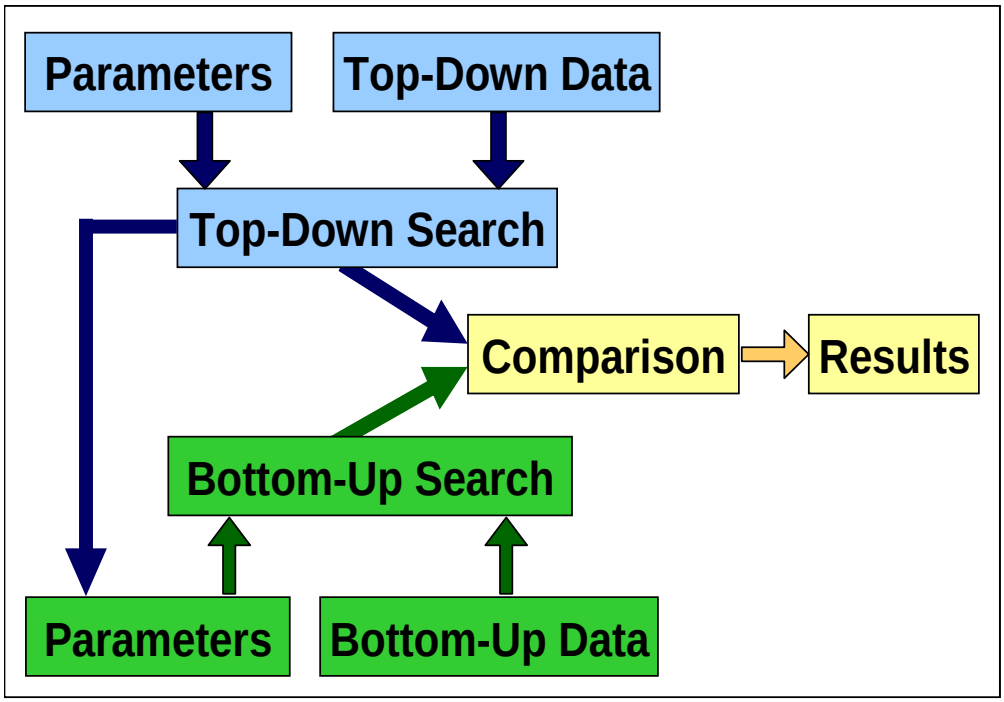


Figure 6.4: Integrated approach of PTMSearch Plus that is able to combine "top-down" and "bottom-up" searching algorithms.

Current "bottom-up" search engines (e.g. Sequest, DBDigger etc.) don't have the possibility to limit the number of PTMs in a single peptide to a reasonable level that could be considered acceptable from a chemical viewpoint [62, 118]. If a peptide has n_X=4 arginines, and arginines can have mono-, di- or tri-methylation (p_X=3), then it results in 256 peptide candidates each with different PTMs. Generally, peptides with 0, 1 or 2 PTMs could be found. For this reason, it seemed to be practical to let the user limit the number of possible PTMs on a single peptide (m) in the software to reduce the searching time. Using the example above, (n_X=4, p_X=3) the number of the different peptide candidates is 13 or 67, using m=1 or m=2, respectively. This simple example clearly demonstrates that the number of PTM peptide candidates, and the time necessary to search them against the experimental "bottom-up" data, can be (drastically) reduced by limiting the number of PTMs that a single peptide can have based on simple chemical viewpoint and our experimental experience. The reduction in the number of peptide candidates, by limiting the maximum PTM/peptide, is even more drastic when different PTMs are assigned to different amino acids (e.g. methylations to arginine and lysines, β-methythiolation to aspartic acid, etc.). Bottom-up inputs are also available for Sequest [62], DBDigger[118], and DTASelect[61] text files, however, the advantages of "top-down" predicted "bottom-up" search cannot be utilized in these cases.

Methods

All proteins, salts, and buffers were obtained from Sigma Chemical Co. (St. Louis, MO). Sequencing grade trypsin was purchased from Promega (Madison, WI). Formic acid was obtained from EM Science (affiliate of Merck KGaA, Darmstadt, Germany). HPLC-grade acetonitrile and water were used for all LC-MS-MS analyses

(Burdick and Jackson, Muskegon, MI). Ultrapure 18 MΩ water used for sample buffers was obtained from Millipore Milli-Q system (Bedford, MA). Fused silica capillary tubing was purchased from Polymicro Technologies (Phoenix, AZ).

Preparation of Protein Standard Mixture and Rhodopseudomonas palustris Ribosomal Proteins

In this study, all prepared samples were divided into two portions. One portion was examined by 1D LC-MS-MS bottom-up mass spectrometry and the other portion of the sample was examined using LC-FT-ICR-MS for top-down mass spectrometry. By correlating the two data sets, using PTMSearch Plus with the same sample, it was possible to identify the proteins, but also to characterize PTMs on the proteins.

Five proteins were used in a five protein mixture: ubiquitin (MW 8 kDa), chicken lysozyme C (MW 14 kDa), bovine ribonuclease A (MW 13 kDa), bovine carbonic anhydrase II (MW 29 kDa), and bovine beta lactoglobulin-B (MW 18 kDa). The proteins were dissolved in HPLC grade water to give a final concentration of 1 mg/mL of each protein, and diluted as required for the analysis. The PSM mixture was digested for bottom-up analysis with sequencing grade trypsin added at 1:20 (wt/wt) of enzyme to protein. The digestions were run with gentle shaking at 37 °C for 12 hours. Samples were immediately desalted with an Omics 100 µl solid phase extraction pipette tip (Phenomenex, Torrance, CA). All samples were frozen at -80°C until LC-MS/MS analysis.

70S ribosomes from *R. palustris* were purified and fractionated using a high salt sucrose cushion and sucrose density fractionation as previously described [119]. For bottom-up analysis acid extracted [120] ribosomal proteins were denatured and reduced

in 6M guanidine HCl, 50 mM Tris-HCl (pH 7.6), with 10 mM DTT at 60 °C for 45 minutes. Afterward, the proteins were digested with 1 μg trypsin overnight at 37 °C. Remaining disulfides were reduced with 10 mM DTT at 60 °C for 45 minutes. To perform top-down analysis the ribosomal samples were neither reduced nor digested.

Results and Discussion

Protein Standard Mixture

A five protein standard mixture consisting of ubiquitin, lysozyme, ribonuclease A, β-lactoglobulin B, and carbonic anhydrase was evaluated with PTMSearch Plus. The protein standard mixture served as a training set to evaluate the performance of the program with an initial simple mixture. To begin the search of the five protein standard, using PTMSearch Plus, a combined top-down and bottom-up search was selected using both the built in simple bottom-up searching method, as well as the top-down and external search option using the DBDigger program as described above [118]. Both programs were used in order to validate the simple built in bottom-up searching method with a known external bottom-up search algorithm to ensure both data sets corresponded.

A directory comprising text files from the top-down data obtained from scripting methods by selecting data rich regions across the total ion chromatogram, as well as a directory of the MS2 files generated from the raw bottom-up MS/MS data files, was input for the five protein search. These data directories can be selected from browser tab at the input sites within the software main screen. Once the appropriate data directories were loaded, a list of PTMs to be searched was input. Within all searches of the protein standard mixture, the only specified PTMs were disulfide bonds and methionine truncation. These two PTMs were selected due to the intact proteins used containing

these modifications. A feature of the PTM function within PTMSearch Plus is the ability to perform certain smart PTM searches (Figure 6.3). For example, the program looks for the number of cystines within a protein and will not allow for more disulfide bonds to be formed than there are cystines to support.

All searching was performed with a database composed of the five proteins, as well as common contaminants to give a total of 43 proteins within the database. Top-down specifications included a maximum mass difference of one dalton and a minimum of three peaks within the isotopic package. Also used in the top-down search specifications were all three FFT data sizes (128, 256, 1024K), and apodization both on and off. For the bottom-up search parameters, a peptide can have a maximum of two missed cleavages, a maximum of two PTMs on a peptide, a minimum of 5 amino acids within a peptide, and a minimum mass of 400 Da and maximum mass of 6000 Da. Of the five proteins searched for within the mixture, four could be identified both from the intact protein data, as well as having supporting peptide data from both bottom-up search types. The identified four proteins with corresponding top-down and bottom-up data included ubiquitin, lysozyme, ribonuclease A, and β-lactoglobulin-B. Carbonic anhydrase is a 29 kDA protein that is difficult to elute from the C4 reverse phase column used in the top-down analysis. Therefore, carbonic anhydrase was identified in the bottom-up searching but not in the top-down data.

Post translational modifications in the form of disulfide bonds were identified on three of the proteins in the integrated search. These included lysozyme with two disulfide bonds, β-lactoglobulin B with two disulfides, as well as ribonuclease A with multiple isoforms and disulfide bonds. Also identified was ubiquitin with a methionine

truncation. All of these modifications were expected due to the use of purchased protein stocks with these known modifications. These modifications were previously known, however, PTMSearch Plus was able to identify them without any prior inputs into the software indicating their presence.

Rhodopseudomonas palustris Ribosomal Proteins

In a recent study by Strader et al. top-down and bottom-up characterization of the ribosome from *R. palustris* was performed. In this study 53 of the 54 orthologs to the *E. coli* ribosomal proteins were identified by bottom-up analysis, and 42 intact protein identifications were obtained by the top-down approach [54]. Following top-down mass measurement, the authors used a manually created intact protein look-up table that contained intact molecular masses, methionine truncated molecular masses, and all possible combinations of methionine truncation with single acetylation and multiple methylations, up to 9, for the entire suit of 54 possible ribosomal proteins. After bottom-up measurement, the authors used SEQUEST [62] to identify peptides with no modifications. Next, numerous single searches of the data using SEQUEST with individual possible PTMs was performed. Once the identifications from both the top-down and bottom-up approach were generated, they were then compared to one another manually to provide conformation of both methods. This manually inspected data set yielded a test set for PTMSearch Plus to test the program with a complex mixture. Since PTMSearch Plus is able to take the top-down intact protein data and the bottom-up MS/MS data and combine them into one single search, thereby eliminating the time consuming manual search and conversion of data, the time to search both raw data sets took only minutes as compared to months for the manual conversion.

When performing the searches with PTMSearch Plus, a combined top-down and bottom-up search was performed, using both the built in simple bottom-up searching, as well as the top-down and external search option using the DBDigger program. Top-down search specifications included a maximum mass difference of one dalton and a minimum of three peaks within the isotopic package. Also used in the top-down search specifications was all three FFT data sizes (128, 256, 1024K), and apodization both on and off. For the bottom-up search parameters, a peptide can have a maximum of two missed cleavages, a maximum of two PTMs on a peptide, a minimum of 5 amino acids within a peptide, and a minimum mass of 400 Da and maximum mass of 6000 Da. Within all searches of the complex ribosomal mixture, the only specified PTMs were methylations, acetylations, and methionine truncation. These PTMs were selected due to the initial study using these modifications, therefore, our search results could be directly compared to the manual results published in the study.

Using PTMSearch Plus, we were able to identify all of the 53 identified by bottom-up analysis, and 42 intact protein identified by the top-down approach within the Strader et al. study [54]. PTMSearch Plus is able to output the proteins identified by bottom-up only, top-down only, and a list of measured intact proteins, with or without PTMs, that have confirming bottom-up data. The examine matches interactive output selection, loads a graphical display showing the identified protein with its corresponding identified peptide and PTMs. This function is able to filter the results with a number of different options to show only the top-down matches that have confirmation with bottom-up data, delete duplicate proteins with the same PTMs, and delete duplicate peptides that confirm the protein plus PTM. Once the results have been filtered, the user is able to view

the identified proteins with their associated PTMs and peptides. Within the results view, the protein name is shown, sequence of the peptides, the PTMs on the protein and the peptides, as well as the b and y ion labeled ms/ms spectra of the peptide. These results interfaces allow for the quick and easy viewing of results. When looking at the proteins that contain supporting peptide data we find a total of 41 *R. palustris* ribosomal proteins. This is consistent with the individual top-down and bottom-up data due to there only being 42 identified top-down peaks to match peptides. The one unidentified top-down peak is L36 where bottom-up could not provide any supporting peptides. These results are consistent with what was seen in the Strader *et al*. study [54]. Figure 6.4 shows the top-down and bottom-up data for ribosomal protein L33 as out put by PTMSearch Plus. The L33 protein was identified with a methylation of the peptide AK*AVTIKIK by bottom-up analysis. The mass of L33 with a methionine truncation and a methylation was identified in the top-down searching. When PTMSearch Plus output the proteins that have confirming peptide data L33 was shown [Figure 6.5].

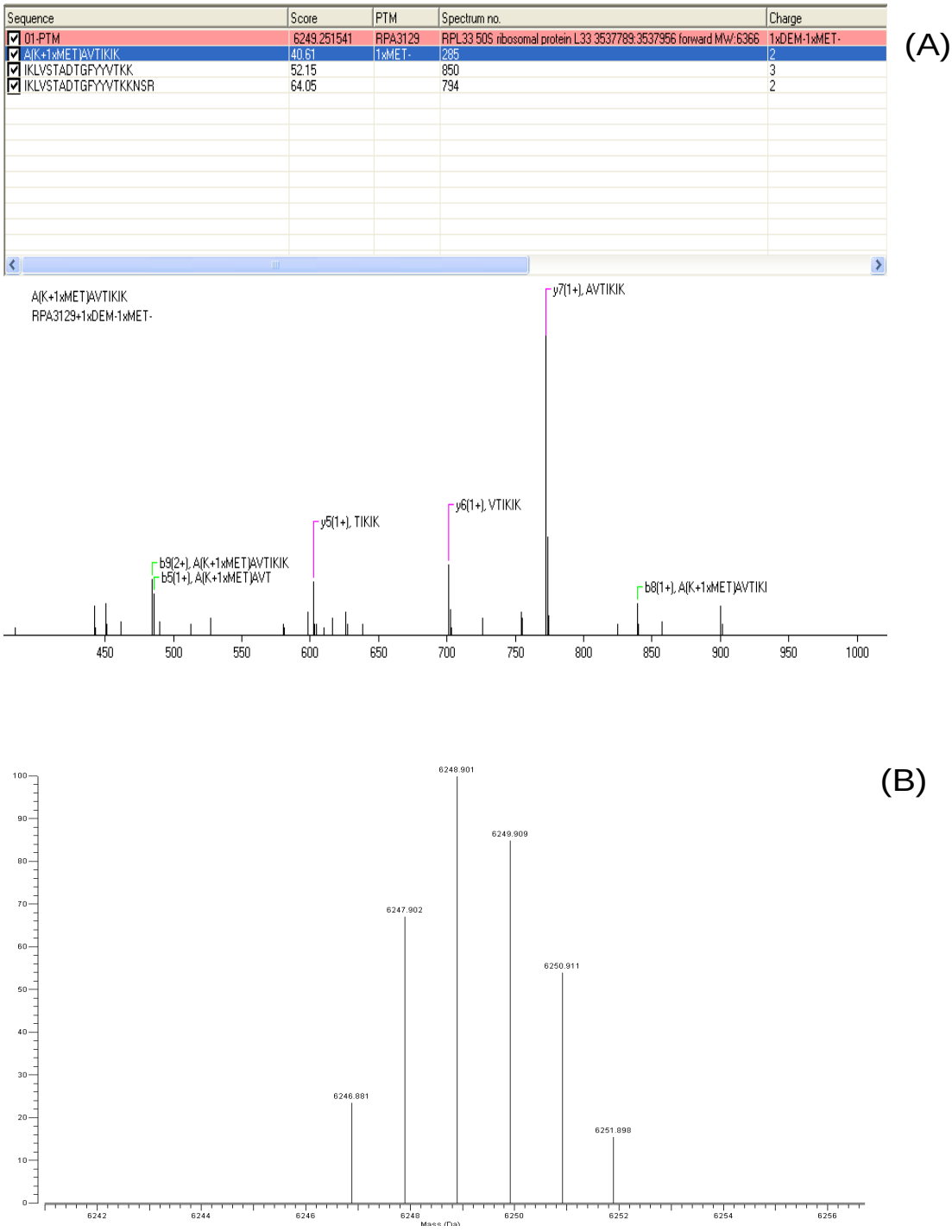


Figure 6.5: Integrated top-down and bottom-up results for the *R. palustris* L33 protein. (A) Shown are the results lists from PTMSearch Plus with the labeled confirming MS/MS spectrum from the bottom-up analysis. The sequence of the peptide is given with the methylation labeled as MET within the sequence. (B) Shown is the top-down spectrum of the intact protein with the de-methionation and a methylation.

Conclusions

PTMSearch Plus provides a novel software for the integration of top-down and bottom-up protein and PTM identification. The software allows for the use of multiple data and instrument platforms to be combined. This methodology provides an integrated top-down and bottom-up searching algorithm that is not only fast but accurate. The software was demonstrated with a protein standard mixture and complex ribosomal protein mixture. All proteins from the protein standard mixture, which was used as a training set, could be identified using PTMSearch Plus. The *R. palustris* complex ribosomal mixture was previously examined in an integrated fashion by manual comparison. Using PTMSearch Plus all of the identified ribosomal proteins identified in the previous study were identified in a fraction of the time. Both of these test cases showed the power of the integrated approach, as well as demonstrating the accuracy and speed of PTMSearch Plus.

Chapter 7

Identification of PTMs and Isoforms from the Versatile Microbe

Rhodopseudomonas palustris Under Three Metabolic States

All of the data presented below are in preparation for submission Heather M. Connelly, Dale A. Pelletier, Vilmos Kertesz, Melissa Thompson, W. Judson Hervey, Tse-Yuan Lu, Patricia K. Lankford, Gregory B. Hurst, Frank W. Larimer, and Robert L. Hettich Top-down Characterization of the Versatile Rhodopseudomonas palustris Microbe Under Three Growth Conditions to Identify PTMs and Isoforms. *Journal of Proteome Research* (2006). *Judson Hervey a graduate student in the genome science and technology program provided signal peptide database. All MS sample preparation, experiments and data analysis were performed by Heather M. Connelly.*

Introduction

Rhodopseudomonas palustris belongs to the α-proteobacteria, and is a purple nonsulfur anoxygenic phototrophic bacterium found in diverse environments from fresh water to soil. One of the unique features of *R. palustris* is its ability to grow and function under many metabolic states. These states include: photoheterotrophic where energy is obtained from light and carbon from organic carbon sources, photoautotrophic where energy is from light and the main source of carbon is from carbon dioxide, chemoheterotrophic where carbon and energy are from organic compounds, and finally chemoautotrophic where energy is from inorganic compounds and carbon from carbon dioxide [10, 11, 12, 13] (Figure 7.1). *R. palustris* has the ability to be a biofuel producer by producing hydrogen gas as a byproduct of nitrogen fixation, as well as a greenhouse gas sink by converting carbon dioxide into cell mass. Since most of these metabolic states can easily be attained in laboratory settings, *R. palustris* is an ideal model system for the study of diverse metabolic modes and their control within a single organism.

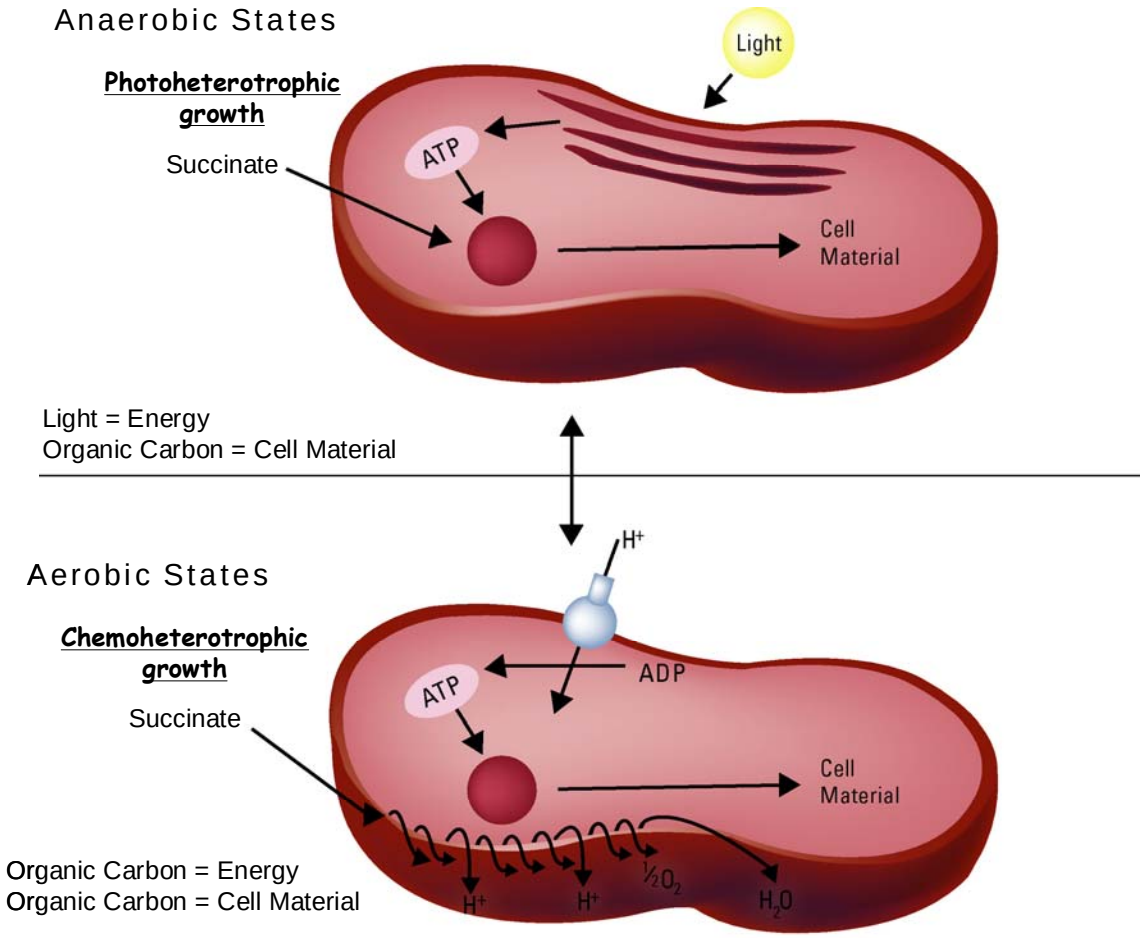


Figure 7.1: Graphical representation of the core metabolic states of *Rhodopseudomonas palustris* interrogated in this study. The top figure illustrates the basic anaerobic state for photoheterotrophic growth in light without oxygen. The bottom figure illustrates the basic aerobic state for chemoheterotrophic growth in the dark with oxygen present. The circle in the center of the cell represents central metabolism. This figure was adapted from Larimer et al. Nat. Biotechnol.2004, 22, 55-60.

Recently, *R. palustris* has been sequenced revealing a 5.4 Mb genome with 4836 potential protein encoding regions [13]. This sequencing and annotation effort other along with proteome profiling, protein-protein interaction studies, global gene knockouts [14], and transcriptome profiling [15], will provide a detailed systems biology characterization of this microbe. A study of the baseline proteome of an *R. palustris* wild-type strain under phototrophic and chemotrophic growth conditions, including variants of each state, was recently completed providing a starting point for understanding this microbe's protein diversity [121].

The goal of this study was to provide the first comprehensive intact protein or "top-down" characterization of *R. palustris*. Intact protein or Top-down mass spectrometry can be used to provide intact protein identification as well as insight into protein modification states, to ascertain the role individual proteins play in the complex metabolism states of *R. palustris*. This powerful method can provide information on the natural state of intact proteins, including details about post-translational modifications (PTM's), truncations, mutations, signal peptides, and isoforms due to top-down mass spectrometry's ability to measure the molecular weight of a protein very accurately and detect any covalent modifications that alters the mass of a protein.

The more common peptide or "bottom-up" mass spectrometric approach involves enzymatic digestion of intact proteins with a protease to generate a peptide mixture. However, bottom-up methods provide a comprehensive list of proteins, vital information about post translational modifications may be missed if the peptides containing the particular modification escape detection. Furthermore, identifying peptides that come from a complex protein mixture may not provide information on the presence of different

isoforms (variations of a protein that may include different states of PTMs) that may exist for a particular protein.

An integrated top-down and bottom-up approach allows for a more complete characterization of proteins due to the unique strengths of each technique. In an integrated approach, intact protein masses from the top-down analysis corresponding to a particular PTM or isoform are compared to the comprehensive list of proteins provided by the bottom-up analysis. This correlation between the two methods can provide information on PTM location and identity, as well as verifying gene start sites within the genome annotation with more certainty. The comprehensiveness of this technique has been previously demonstrated in studies of the *Shewanella oneidensis* proteome as well as the 70S ribosomal complex from *Rhodopseudomonas palustris* [54, 35]. Combing the strengths of two important mass spectrometric techniques such as "top-down" and "bottom-up", proteomic strategies provides a powerful tool to examine proteins from selected growth states of *R. palustris*.

Three growth states of *R. palustris* were interrogated with this integrated top-down and bottom-up approach. These three growth states consist of two categories: aerobic growth in the dark (chemotrophic) and anaerobic growth in light (phototrophic). The main growth state was the anaerobic photoheterotrophic growth mode, with light providing the energy, organic carbon in the form of succinate providing the carbon source for cell material, and ammonia serving as the nitrogen source. The second growth state examined was a variant of the photoheterotrophic growth mode in which nitrogen fixation is performed. In this state, nitrogen gas was substituted for ammonia as the nitrogen source forcing the cells to fix nitrogen. The final growth state examined was the

aerobic state or chemoheterotrophic growth state, in which cells were grown aerobically in the dark, with succinate as both the carbon and energy source, and nitrogen serving as the ammonia source. These three growth conditions provided an opportunity to examine both intact proteins and how post translational modifications (PTMs) from *R. palustris* play a role in the complex metabolic processes carried out by this organism.

This study provides the first large-scale characterization of these three growth states of *R. palustris* by an integrated top-down and bottom-up approach. This global measurement strategy can provide information on intact proteins, including PTMs, isoforms, and signal peptides from a given growth state. This technological approach provides information on the function and location of proteins, as well as providing confirming peptide MS/MS data. This tool is especially powerful when determining what modification states play a role in the switch between different growth conditions, characterizing known and unknown proteins, and determining trends within protein expression across the chosen metabolic states.

Material and Methods

Chemicals and Reagents

All salts, buffers, dithiothreitol (DTT), guanidine HCl, trifluoroacetic acid, phenyl methyl sulfonyl fluoride (PMSF), were obtained from Sigma Chemical Co. (St. Louis, MO). Sequencing-grade trypsin was purchased from Promega (Madison, WI). Formic acid was obtained from EM Science (Affiliate of Merck KgaA, Darmstadt, Germany). HPLC grade acetonitrile and water were used for all LC-MS analyses (Burdick & Jackson, Muskegon, MI). Ultrapure 18 MΩ water used for sample buffers was obtained from a

Millipore Milli-Q system (Bedford, MA). Fused silica capillary tubing was purchased from Polymicro Technologies (Phoenix, AZ).

Cell Growth and Protein Fractionation

R. palustris strain CGA010, a hydrogen-utilizing derivative of the sequenced strain (unpublished C. S. Harwood) and referred to here as the wild-type strain, was grown under the three conditions outlined in the Introduction section. Wild type *R. palustris* cells were grown anaerobically in light or aerobically in dark on defined mineral medium at 30 °C to mid-log phase (OD 660 nm = 0.6). Carbon sources were added to a final concentration of 10 mM succinate, 10 mM sodium bicarbonate. For the photoheterotrophic N_2 fixing cultures, ammonium sulfate was replaced by sodium sulfate in the culture medium and N_2 gas was supplied in the headspace. Chemoheterotrophic cells were grown aerobically in the dark with shaking at 200 rpm; phototrophic cells were grown anaerobically in the light with mixing with a stir bar. All anaerobic cultures were illuminated with 40 or 60 W incandescent light bulbs from multiple directions. 4-5 liters of cells were grown for all three states and pooled together for each state. The cell pellet, obtained by centrifugation at 1000 X g for 10 minutes, from each growth state was French Pressed to yield 60-120 mg of protein for each of the three growth states. Cell extract was centrifuged at 10,000g for 35 minutes in a Sorvall centrifuge to remove all unbroken cells. Protein extract was used for off-line anion exchange FPLC fractionation. To perform off-line anion exchange chromatography 60 mg of protein was injected onto a 5 ml HiTrap (HiTrap SP HP, Amersham Pharmacia) ion exchange column connected to an AKTA (Amersham Pharmacia) FPLC system. After protein injection, a 30 minute ammonium acetate gradient was run from 0.2 M to 2 M. Twenty fractions from each

growth state (total of 60 from 3 growth states) were determined to have sufficient protein concentrations (400 μg) by a Bradford protein assay. Each FPLC fraction obtained was then divided into two equal protein concentration portions. One portion was examined by 1D LC-MS-MS bottom-up mass spectrometry and the other portion of the sample was examined using LC-FTICR-MS for top-down mass spectrometry. All fractions were analyzed with 1X coverage with top-down methods due to the proteins precipitating upon freezing. Bottom-up analysis was performed with 2X coverage on all fractions.

Data Analysis

All resulting top-down and bottom-up data sets were analyzed with two methods. In the first method, the SEQUEST algorithm was used to identify MS-MS spectra with their counterparts predicted from a protein sequence database [62]. For all database searches, an *R. palustris* proteome database was used, which contained 4,833 proteins and 36 common contaminants. All resultant output files from SEQUEST were filtered by DTASelect [61] at the 1-peptide, 2-peptides and 3-peptides level with the following parameters: SEQUEST, delCN of at least 0.08 and cross-correlation scores (Xcorrs) of at least 1.8 (+1), 2.5 (+2) and 3.5 (+3). Once filtered the results were analyzed by Contrast [61] for comparison. In the second method, integrated top-down and bottom-up searching was performed with the PTMSearch Plus software developed at Oak Ridge National Laboratory. Output files containing bottom-up data from PTMsearch Plus were filtered by DTASelect [61] at the 2-peptides level with the following parameters: MASPIC [118], scores of at least 23 (+1), 28 (+2) and 43 (+3). The output files containing top-down data were filtered with at least three peaks within the isotopic package, a 3000 Da mass cutoff and a relative abundance of at least 10%. The

PTMSearch Plus program allows for the combined searching of both the top-down and bottom-up data sets, as well as allowing for the searching of a defined set of PTMs. In the integrated top-down and bottom-up data searches a standard set of PTMs were searched for including: methylation, acetylation, de-methionation, and disulfide bonds (restricted to top-down data). Less common PTMs such as uridylylation were searched individually. All data outputs generated were manually inspected and then compared using Microsoft Access (Microsoft Corp., Redmond, WA).

Results and Discussion

Approximately twenty fractions were obtained from the off line FPLC separation of the protein lysate from each of the three anaerobic, nitrogen fixing, and aerobic growth states. Off line FPLC was used to separate the large complex mixtures of proteins from the three growth states for top-down analysis, due to its proven ability to reduce down the complexity of the mixture. Therefore, by reducing the complexity of the protein mixture, this method allowed for better separation from the on-line HPLC methods employed, as well as more comprehensive protein identifications. Top-down methodologies are a powerful tool, but some limitations do exist such as on-line chromatography of intact proteins is difficult due to the wide range of protein sizes and hydrophobicities within the complex mixtures used. This method of off line FPLC fractionation followed by on line HPLC takes a large amount of protein starting material (in the milligram range), although, this is not of great concern due to the ability to produce more than enough material (~120 mg) from the 4 liters of culture from each growth state. Another area of concern using this strategy is the loss of protein during the off line separation. This problem is unavoidable, due to the need to have a prior separation of the complex protein

mixture before the top-down analysis. By providing this initial separation step with FPLC, we increase the overall ability to analyze and identify more proteins than with no initial separation. We have found this technique of off line fractionation followed by on line HPLC, to be highly reproducible and simple to implement for a large-scale study of multiple samples [35].

Each data set generated with the integrated top-down bottom-up approach was searched with PTMSearch Plus, which combined the top-down and bottom-up data set searches to provide positive identifications of both proteins and their associated PTMs. PTMSearch Plus is a new search algorithm developed at Oak Ridge National Laboratory (ORNL), which provides the first integrated top-down and bottom-up searching algorithm that allows the user to select the number and types of PTMs they wish to search for. As well as the integrated searching approach, all bottom-up MS/MS data sets were searched with SEQUEST [62], filtered with DTASelect [61], and compared with Contrast [61]. Since SEQUEST is a proven search tool within the community; the search results from SEQUEST were used to manually verify the outputs from PTMSearch Plus were accurate. The results from the integrated top-down and bottom-up searches for all three growth states allowing up to 10 methylations, 2 acetylations, N-terminal methionine truncation, and disulfide bonds on the intact proteins are shown in Table 7.1. These PTMs and amounts were used for searching due to biological constrains and to keep the search space and results manageable. For example, only one N-terminal methionine truncation can be present on a protein and the number of disulfide bonds is restricted to the number of cysteines present within the protein sequence. Also shown in Table 7.1 are the results for the SEQUEST searches of the bottom-up MS/MS data with no specified

Table 7.1: Number of identified proteins from all three searching methods.

Growth Condition	PTMSearch Plus TDBU PTM [a]	PTMSearch Plus BU PTMs [b]	SEQUEST BU no PTM [c]
Anaerobic	119	853	465
Nitrogen Fixing	214	785	295
Aerobic	426	1373	512
Total Non-Redundant Proteins Identified	599	1908	713

[a] **Top-down and bottom-up searching with PTMs performed with PTMSearch Plus**
[b] **Bottom-up searching with PTMs performed with PTMSearch Plus**
[c] **Bottom-up searching performed with SEQUEST and no PTMS**

PTMs and the PTMSearch Plus bottom-up results allowing for the PTMs listed above. There were more proteins identified using the bottom-up searching containing PTMs. The identification of more proteins with the PTMSearch Plus program is due to the ability to confidently identify MS/MS spectra from a peptide containing PTMs. These MS/MS spectrum would have been unidentified in the SEQUEST program. This is due to the inability of SEQUEST to perform searches with multiple modifications specified. Therefore, using PTMSearch Plus provides more comprehensive identifications for the data set. One method used to ensure our searches were identifying proteins correctly was to examine some of the common proteins that one would expect to find, such as elongation factors, chaperonin GroES, and nitrogen regulatory proteins. Table 7.2 provides a list of these expected proteins found in both the top-down and bottom-up data sets. The percent sequence coverage of the protein from the bottom-up data is provided in Table 7.2 as well as the ppm error from the top-down data.

Both the top-down and bottom-up data sets were evaluated separately and in an integrated approach within this study. Tthe bottom-up method provides a confident list of proteins, using MS/MS data, but there are instances where bottom-up is unable to provide identifications. These proteins missed with the bottom-up method are generally very amenable to the top-down approach, due to the proteins being within a size range that works well for top-down measurements (3-15 kDa). Generally, the unidentifiable proteins from bottom-up are small in size or have very few tryptic sites. Small proteins when digested form peptides that are too small to be seen within the mass spectrometer (less than 400 Da).

Table 7.2: Expected proteins and their percent sequence coverage and mass accuracy.

Gene Number	Category	Product	% Sequence Coverage	Mass Accuracy
RPA3053	Transcription	Cold Shock Protein	73.1	1.2
RPA3672	Transcription	Cold Shock Protein	52.4	12
RPA2513	Translation	Elongation Factor P	41	28.1
RPA1141	Cellular Processes	Chaperonin GroES1	49	11.1
RPA2165	Cellular Processes	Chaperonin GroES2	58.7	1.7
RPA2966	Signal Transduction	GlnB Nitrogen Regulatory Protein PII	51.8	-4.1

Also, proteins with few tryptic sites generate peptides that are to large to be measured with the mass spectrometers employed in bottom-up. Table 7.3 provides examples of some of these proteins, where the tryptic peptides used are outside the 400-6000 Da range generally seen in the bottom-up method, but were identified by top-down. Also, within this study, positive protein identifications from bottom-up searching require two unique peptides. In the cases presented in Table 7.3, some of these proteins do not have the required two unique peptides for a positive identification. Therefore, top-down methods alone are able to add another level of information above the identification of PTMs and isoforms generally considered. Contained within the top-down data sets are proteins with good isotopic resolution and mass accuracy, but are unidentified. These unidentified proteins may be from degradation and truncation products or the result of missed start sites within the protein annotation process.

Due to the focus of this research being to confidently identify intact proteins and their associated PTMs, all data we will focus our biological analysis on will be from the integrated TDBU data obtained for the PTMSearch Plus program. From the integrated Top-down and bottom-up searching, a total of 599 non-redundant proteins were identified from all three growth states [Table 7.4]. Table 7.4 shows all 599 proteins identified, organized by functional categories. These 599 proteins include both proteins identified with and without PTMs and all have bottom-up MS/MS confirmation.

Table 7.3: Proteins not identified by bottom-up analysis that were identified by top-down.

Protein	Intact Mass	Peptides	Mass Peptide
RPA0214	8651.6691	1-MALGEPQEVPNDPGPVTPPPEVPPSTPGTPTEPPLEQPPGN PNPDIPPPEEPGAPPQPNELPGQMPAEVPMQSPG**R**-77	7929.859
		78-SVPNPGVA-85	739.827
RPA1952	7593.1835	1-TAELNILGVFVPTILICAAAAFILTSLVS**R**-30	3117.788
		31-LLVWLNFYHLVWHHTLFNLTIFVVIVFVALGLVSGWPQ-68	4493.411
RPA1773	11366.5634	1-M**K**-2	277.388
		3-WLYLLIAIVAEVVGTSAL**K**-21	2059.521
		22-ASQGFTVLLPSVLVVVGYGAAFYFLSLTLSSISVGIAYA WSGIGIVLISAVGWLWFGQALDTAAIIGIAFIIAGVGIINFFSNVSAH-109	9065.685
RPA1085	14242.7461	1-M**K**-2	277.388
		3-YAGILAAFALGASVAGADAGSLVYTPTNPAFGGS PLNGSWQMQQATAGNHFN**R**-55	5297.859
		56-AAPTSGPQQLTQSQIFAQQLQSQLYASLANQVT QAIFGANAQQSGTFSFQGTTISFA**K**-113	6108.735
		114-VDGQTNITINDGSTVTQISLPTVTH-138	2611.847

Table 7.4: All 599 proteins identified from the three growth states of *R. palustris*

Gene Number	Aerobic	Anaerobic	Nitrogen Fixing	PTMs	BU Seq. Cov.[e]	Protein Annotation
RPA0008	10127.686[a], S	10086.001, W[d]		1xDEM-3xMET 1xDEM	30	circadian clock protein
RPA0010	18760.967, S[b]			4xMET		transcriptional regulator, probable glutamate
RPA0036	17908.472, W					conserved unknown protein
RPA0038		13314.382, M[c]		1xMET	70	ribosomal protein L20
RPA0039		7457.667, S		1xDEM-3xMET	27	50S ribosomal protein L35
RPA0040	21979.05, S			4xMET	21	translation initiation factor IF-3
RPA0052		16447.565, M		1xDEM-4xMET		putative nitrogen regulatory IIA protein(enzyme
RPA0054			15894.144, W	1xDEM-3xMET	12	putative small heat shock protein
RPA0059		43156.031, M		1xDEM-1xMET	51	L-carnitine dehydratase/bile acid-inducible
RPA0090	15149.266, W			1xDEM	82	hypothetical protein
RPA0092	10019.455, W			1xDEM-2xMET	50	conserved hypothetical protein
RPA0155		16656.413, M		6xMET	10	putative tolR/exbD protein
RPA0158	13359.566, S		13359.564, M	1xDEM 1xDEM	42	putative ribosomal protein L21
RPA0159	9477.775, S		9463.792, S	1xDEM-2xMET 1xDEM-1xMET	40	ribosomal protein L27
RPA0160			22277.636, S	1xDEM-3xMET	9	possible acetyltransferases.
RPA0177	32018.782, W			6xMET	9	putative H+-transporting ATP synthase gamma
RPA0179	19541.781, W			2xMET	13	putative H+-transporting ATP synthase delta
RPA0203	20990.138, M			1xDEM-3xMET	78	heme exporter protein A (heme ABC transporter
RPA0207	13656.301, S		13699.846, M	1xDEM-3xMET 1xDEM-6xMET	28	unknown protein
RPA0222		19744.263, M		1xDEM-8xMET	61	Beta-Ig-H3/Fasciclin domain
RPA0224	28876.763, S			1xDIS-6xMET	9	similar to eukaryotic molybdopterin
RPA0233	31223.309, S			1xDEM-3xMET	73	putative Citrate lyase beta chain (acyl lyase
RPA0235	22365.940, M	22366.559, M	22366.103, S	3xMET 3xMET 3xMET	15	3-isopropylmalate dehydratase small subunit

Table 7.4: Continued

Gene Number	Aerobic	Anaerobic	Nitrogen Fixing	PTMs	BU Seq. Cov.[e]	Protein Annotation
RPA0241	14297.711, S	14297.834, S	14297.470, S		32	50s ribosomal protein L19
RPA0244	12017.821, S			1xDEM	14	ribosomal protein S16
			12017.633, S	1xDEM		
RPA0246	13472.790, W				14	PilT protein, N-terminal
RPA0263	15891.873, M			1xDEM-4xMET	27	Protein of unknown function UPF0047
RPA0267	32750.676, M			1xDEM-5xMET	4	possible thioredoxin
RPA0272		12360.825, M			17	GlnK, nitrogen regulatory protein P-II
RPA0274			12367.542, M	1xURY	13	GlnK, nitrogen regulatory protein P-II
RPA0276	12559.245, S			2-4xMET	17	PAP/25A core domain:DNA polymerase, beta-like
RPA0282	16018.135, W			1xDEM	8	possible transcriptional regulator
RPA0283	24983.847, S			1xDEM	9	putative two-component response regulator
RPA0285	23538.857, W			1xDIS-9xMET	6	Protein of unknown function UPF0001
RPA0292	30801.005, W			1xDEM-1xDIS-3xMET	6	chromosome partitioning protein, ParA
RPA0298	30818.180, W			1xDEM-6xMET	17	DUF299
RPA0301	25755.166, W			1xDEM-1xDIS-4xMET	5	putative DNA polymerase III epsilon chain
RPA0311	23669.290, W			1xDIS-3xMET	8	imidazoleglycerol-phosphate synthase,
		23669.449, M		1xDIS-3xMET		
RPA0323	14372.435, M			8xMET	7	Protein of unknown function UPF0102
RPA0326	13489.857, W			1xDEM-9xMET	12	DUF24, predicted transcriptional regulator,
			13490.434, M	1xDEM-9xMET		
RPA0329	26017.808, S			2xDIS-3xMET	4	ribonuclease PH
RPA0331	22173.571, M			1xDEM-7xMET	15	possible heat shock protein
			22305.627, M	7xMET		
RPA0335	22366.393, M			1xDEM-8xMET	9	putative phospholipid N-methyltransferase
			22366.586, M	1xDEM-8xMET		
RPA0350	16176.801, S			9xMET	13	putative patch repair protein
		16176.339, S		9xMET		
			16119.958, M	5xMET		
RPA0354	11347.234, S			1xDEM-5xMET	22	putative pts system phosphocarrier protein HPr
RPA0356	15924.889, M			3xMET	9	conserved hypothetical protein
RPA0359	26896.572, S			5xMET	5	conserved unknown protein
		26866.940, W		3xMET		

Table 7.4: Continued

Gene Number	Aerobic	Anaerobic	Nitrogen Fixing	PTMs	BU Seq. Cov.[e]	Protein Annotation
RPA0366	12741.565, S			1xDEM-3-4xMET	11	unknown protein
RPA0373	16027.073, W			2xDIS-8xMET	13	thioredoxin
			16027.527, W	2xDIS-8xMET		
RPA0377	30252.460, W			1xDIS-7xMET	4	conserved unknown protein
RPA0384	20438.524, W			1xDIS-8xMET	8	conserved unknown protein
RPA0395	21641.686, M			1xDIS-2xMET	10	Metal dependent phosphohydrolase, HD region
RPA0403	14135.043, W			9xMET	9	conserved hypothetical protein
RPA0414		11855.153, M		4xMET	16	DUF167
			11870.371, M	5xMET		
RPA0433	10011.569, S			1xDEM	28	ribosomal protein S15
		10011.231, M		1xDEM		
			10011.035, M	1xDEM		
RPA0435			15379.549, S	1xDEM-2xMET	12	putative ribosome-binding factor A
RPA0443	15004.785, W			1xDEM-3xMET	10	possible transcriptional regulator
RPA0450	16585.565, M			1xDEM-1xDIS	20	ferric uptake regulation protein
			16644.153, S	1xDEM-4xMET		
RPA0453			20348.108, M	4xMET	57	possible NifU-like domain (residues 119-187)
RPA0489			12526.718, M	1xDIS-1xMET	9	ferredoxin II
RPA0490			8068.031, M	1xDEM-9xMET	25	conserved hypothetical protein
RPA0493	10848.394, S			1xDEM	28	50S ribosomal protein L28
			10876.330, S	1xDEM-2xMET		
RPA0501	9422.580, S			1xDEM	13	BolA-like protein
RPA0511		33493.706, W		10xMET	8	PpiC-type peptidyl-prolyl cis-trans isomerase
RPA0517	17835.864, W			1xDEM-8xMET	15	putative transcriptional regulator (Fur family)
RPA0526						50S ribosomal protein L32
RPA0532		25590.320, M		1xDIS-9xMET	13	beta-ketothiolase, acetoacetyl-CoA reductase
RPA0543	18434.769, S			1xDEM-1xDIS-6xMET	13	unknown protein
RPA0571	20007.852, S			1xDEM	49	two-component, response regulator
			20105.929, S	1xDEM-7xMET		

Table 7.4: Continued

Gene Number	Aerobic	Anaerobic	Nitrogen Fixing	PTMs	BU Seq. Cov.[e]	Protein Annotation
RPA0578	9423.369, S			1xDIS-1xMET	17	unknown protein
RPA0594	14966.448, S			1xDEM-1xDIS-7xMET	17	putative mutator protein mutT
RPA0598	9622.031, W			1xDEM-1xDIS	22	putative glutaredoxin
RPA0600		18099.337, S		1xDEM-7xMET	16	conserved hypothetical protein
RPA0607	30498.225, S			1xDEM-2xDIS-3xMET	7	putative protoporphyrinogen oxidase, hemK
RPA0609			29001.611, S	1xDEM-3xMET	4	conserved hypothetical protein
RPA0616	11188.397, S	11188.089, M	11188.397, M		22	Uncharacterized BCR
RPA0617			21259.868, S	7xMET	9	putative recombination protein recR
RPA0618	14793.278, S			1xDIS-4xMET	10	unknown protein
RPA0626		29733.197, W			6	2,3,4,5-tetrahydropyridine-2-carboxylate
RPA0629		31499.295, W		1xMET	5	putative acetylglutamate kinase
RPA0633			12907.673, M	2xMET	7	probable ribonuclease p protein component
RPA0643	13969.928, W			1xDEM-3xMET	13	conserved hypothetical protein
RPA0646	9413.638, M			1xDEM-4xMET	39	conserved hypothetical protein
RPA0650	27433.140, S			2xDIS-4xMET	8	cyclohex-1-ene-1-carboxyl-CoA hydratase
RPA0653	28669.548, S			2xDIS-7xMET	30	2-ketocyclohexanecarboxyl-CoA hydrolase
RPA0662			8848.975, S	1xDEM-4xDIS-3xMET	19	ferredoxin
RPA0663	17079.605, S			1xMET	60	transcriptional regulator
RPA0673	25777.576, S			1xDEM-9xMET	50	transcriptional activator
RPA0687	23830.531, M			6xMET	40	conserved hypothetical protein
RPA0688	20543.116, M			5xMET	60	ATP-binding component, PhnN protein, possible
RPA0702		43081.468, S		3xMET	30	possible phosphonate ABC transporter, permease
RPA0703			25979.575, S	1xDIS-4xMET	30	conserved hypothetical protein
RPA0704	35909.585, W			7xMET	40	conserved unknown protein

Table 7.4: Continued

Gene Number	Aerobic	Anaerobic	Nitrogen Fixing	PTMs	BU Seq. Cov.[e]	Protein Annotation
RPA0707			11958.916, M	7xMET	11	putative periplasmic divalent cation resistance
RPA0714	18760.967, S			1xDEM-9xMET	70	bifunctional cobinamide kinase, cobinamide
RPA0717			23441.430, S	9xMET	80	putative cob(I)alamin adenosyltransferase
RPA0729	20656.598, M			4xMET	40	conserved hypothetical protein
		20627.654, S		2xMET		
RPA0739	21203.439, S			2xDIS-4xMET	12	putative cytochrome c
RPA0767	34867.931, W			1xDEM-3xMET	40	PAS domain:GGDEF:PAC motif
RPA0771	10002.695, M			4xMET	13	possible protein commonly found in insertion
			9926.698, M	1xDEM-8xMET		
RPA0775	19030.859, M			4xMET	80	hypothetical protein
RPA0791	30019.876, W			1xDEM-2xDIS-7xMET	60	similar to Staphylococcus nuclease (SNase-like)
RPA0795	22822.941, S			1xDEM-8xMET	70	possible SOS-response transcriptional repressor
RPA0830			23441.430, S	5xMET	47	conserved unknown protein
RPA0843	17712.212, M			9xMET	98	putative Fo ATP synthase B chain
RPA0844	19203.863, M			1xMET	12	putative FoF1 ATP synthase, subunit B'
RPA0855		39499.553, M		1xDEM-1xMET	37	Beta-lactamase-like
RPA0866	14264.828, W			1xDEM-6xMET	12	putative nucleoside diphosphate kinase regulator
RPA0868	6011.702, S			1xDEM-5xMET	33	hypothetical protein
RPA0885			10254.021, W	1xDEM-4xMET	10	conserved hypothetical protein
RPA0893	15051.893, W			8xMET	53	conserved hypothetical protein
			14893.115, S	1xDEM-6xMET		
RPA0903	20630.974, S			4xMET	35	putative transcriptional regulator
RPA0907	13450.536, M			8xMET	18	possible response regulator receiver domain
RPA0917			18515.305, M	1xDEM-6xMET	12	Transcriptional Regulator, AraC Family
RPA0918	8566.680, M				20	possible 50S ribosomal protein L31
RPA0920	17135.907, S			5xMET	11	GCN5-related N-acetyltransferase
RPA0927			7731.168, W	1xDEM-1xDIS-4xMET	28	probable transcriptional regulator

Table 7.4: Continued

Gene Number	Aerobic	Anaerobic	Nitrogen Fixing	PTMs	BU Seq. Cov.[e]	Protein Annotation
RPA0930	21414.056, M			1xDEM-1xDIS-5xMET	45	possible 3-octaprenyl-4-hydroxybenzoate
RPA0932	19542.010, S			1xDEM-3xMET	13	conserved unknown protein
RPA0941	11384.043, W			1xDEM-4xMET	21	conserved hypothetical protein
RPA0942	10806.326, W			9xMET	11	conserved hypothetical protein
RPA0953	9142.316, S			1xDEM	26	possible exodeoxyribonuclease small subunit
			9286.551, W	1xMET		
RPA0956			17064.368, W	1xDIS-7xMET	44	hypothetical protein
RPA0973	12572.113, S			2xDIS-3xMET	17	hydrogenase formation/expression protein hypA
RPA0977		41253.600, S		7xMET	32	hydrogenase expression/formation protein hypD
RPA0993	20540.812, W			1xDEM-4xMET	13	possible alpha-ribazole-5`-phosphate phosphatase
RPA0999		22495.358, M		9xMET	54	conserved hypothetical protein
RPA1000	14479.792, M			1xDIS-3xMET	11	Nitrogenase-associated protein:Arsenate
RPA1017			13746.840, M	1xDEM-1xDIS-8xMET	78	Nitrogen fixation-related protein
RPA1019		12257.701, S		1xDEM-1xDIS-5xMET	11	possible transcriptional activator HlyU
RPA1025			13455.728, M	1xDIS-9xMET	74	possible Ectothiorhodospira Vacuolata
RPA1030	26939.380, S				15	possible CoA transferase, subunit B
RPA1061	31284.422, S			2xMET	46	possible polyketide synthesis protein
RPA1064			18078.781, S		60	conserved hypothetical protein
RPA1066	10119.106, W			6xMET	22	hypothetical protein
RPA1088			12590.446, M	1xDEM-8xMET	14	hypothetical protein
RPA1090	23792.666, S			1xDEM-1xMET	15	possible nitrogen regulator
RPA1097	26941.767, S				52	DUF28
RPA1100	21697.417, S			5xMET	73	RuvA; Holliday branch migration protein
RPA1106		18345.992, W		1xDEM-6xMET	11	conserved hypothetical protein

Table 7.4: Continued

Gene Number	Aerobic	Anaerobic	Nitrogen Fixing	PTMs	BU Seq. Cov.[e]	Protein Annotation
RPA1107			19521.468, S	1xDEM-6xMET	90	possible transcriptional regulator
RPA1108	14415.533, W			1xDEM-1xDIS-4xMET	17	Myb DNA-binding domain:DGPF domain
RPA1111	9426.774, W			1xDIS-7xMET	12	hypothetical protein
RPA1141	10493.013, S	10492.701, S	10493.154, S	1xDEM 1xDEM 1xDEM	40	chaperonin GroES1, cpn10
RPA1152		18166.729, M		4xMET	23	hypothetical protein
RPA1157			9553.862, S	8xMET	73	conserved unknown protein
RPA1160	12698.756, W			5xMET	11	conserved unknown protein
RPA1168	17026.930, W		17027.435, M	1xDEM-2xMET 1xDEM-2xMET	15	molybdopterin converting factor, subunit 2
RPA1173	9369.198, S			5xMET	13	possible cold shock protein
RPA1175	14420.772, W			1xDEM-8xMET	24	chemotaxis protein CheY4
RPA1191	17028.760, S			7xMET	73	putative RNA methyltransferase
RPA1228			19751.976, S	1xDEM-9xMET	59	putative 2-oxoglutarate ferredoxin
RPA1263			10369.363, M	6xMET	27	putative II.1 protein
RPA1271	16026.872, M			8xMET	63	conserved hypothetical protein
RPA1278	16179.437, M	16135.775, W		1xDEM-6xMET 1xDEM-3xMET	18	GatB/Yqey
RPA1279			10911.295, M	1xDEM-7xMET	79	hypothetical protein
RPA1289	12010.505, M			1xDEM-6xMET	11	hypothetical protein
RPA1291	11431.413, W			8xMET	32	putative proteic killer suppression protein
RPA1302			12590.586, M	1xDEM-2xMET	90	unknown protein
RPA1333	22674.886, M				73	conserved hypothetical protein
RPA1342			9999.405, M	1xDEM-4xMET	21	hypothetical protein
RPA1344	17028.852, S			6xMET	14	hypothetical protein
RPA1361			7132.965, M	9xMET	24	hypothetical protein
RPA1366	12350.218, S			8xMET	20	putative sulfur oxidation protein
RPA1390	24811.042, M			2xDIS-5xMET	69	conserved hypothetical protein
RPA1392	19630.240, M			1xDEM	77	nitroreductase family proteins
RPA1414	31284.422, S			9xMET	11	MaoC-like dehydratase

Table 7.4: Continued

Gene Number	Aerobic	Anaerobic	Nitrogen Fixing	PTMs	BU Seq. Cov.[e]	Protein Annotation
RPA1416	25863.037, S			1xDIS-9xMET	54	putative branched-chain amino acid transport
RPA1441	28744.987, M			2xMET	77	possible uridylate kinase
		28829.754, S		8xMET		
RPA1442	28899.592, M			7xMET	85	possible uridine monophosphate kinase
			28885.179, S	6xMET		
RPA1454			9368.408, W	1xDEM-4xMET	22	hypothetical protein
RPA1455			16811.588, S	1xDIS-5xMET	40	nitric-oxide reductase subunit C
RPA1475	8778.512, W			9xMET	17	hypothetical protein
RPA1500	21499.481, M			1xDIS-4xMET	11	unknown protein
RPA1535	14599.861, W			1xDEM-7xMET	80	cytochrome c2
			14514.801, M	1xDEM-1xMET		
RPA1551		11037.662, W		1xDEM-8xMET	18	hypothetical protein
			11084.940, W	2xMET		
RPA1578	28900.136, S			8xMET	97	ferredoxin--NADP+ reductase
RPA1586	27968.897, S			1xDEM-1xDIS-1xMET	85	putative short-chain dehydrogenase/reductase
RPA1587	9927.322, M			3xMET	22	hypothetical protein
RPA1589			23442.195, S	1xDEM-1xMET	12	30S ribosomal protein S4
RPA1591			13697.058, M	3xMET	15	conserved unknown protein
RPA1593	18006.889, S			1xDEM-7xMET	50	Thioesterase superfamily
RPA1600	8272.686, S			1xDEM-1xMET	30	BolA-like protein
		8273.949, S		1xDEM-1xMET		
			8301.483, S	1xDEM-3xMET		
RPA1606		14218.167, M		8xMET	79	conserved unknown protein
			14245.615, M	10xMET		
RPA1615	28867.609, M			1xDEM-4xMET	61	putative methyltransferase
RPA1617			21315.981, M	8xMET	52	ErfK/YbiS/YcfS/YnhG
RPA1620			11417.855, M	1xDEM-1xMET	25	unknown protein
RPA1629			13456.672, M	1xDEM-6xMET	58	chemotaxis response regulator
RPA1634	16381.157, M			4xMET	12	conserved unknown protein
RPA1645	8065.034, W			4xMET	33	unknown protein
RPA1659			22884.739, S	1xDEM	25	conserved unknown protein
RPA1661			10306.118, W	1xDIS-3xMET	24	DUF156
RPA1682	22620.915, W			6xMET	72	putative two component response regulator
RPA1693	22557.316, W			3xMET	86	superoxide dismutase
RPA1697	19054.936, S			1xDEM-6xMET	84	Competence-damaged protein
RPA1717			14245.615, S	1xDEM-7xMET	17	hypothetical protein

Table 7.4: Continued

Gene Number	Aerobic	Anaerobic	Nitrogen Fixing	PTMs	BU Seq. Cov.[e]	Protein Annotation
RPA1719	11861.782, W			5xDIS-3xMET	23	Protein of unknown function UPF0153
RPA1726	35298.166, S			1xDEM-1xDIS-6xMET	56	putative oxidoreductase
RPA1757			26980.102, M	1xDEM-1xDIS-7xMET	31	possible oxoacyl carrier protein reductase
RPA1777	15295.070, M			1xDEM-2xMET	11	DUF35
RPA1788	15964.473, M		15894.144, M	1xDIS-8xMET 1xDIS-3xMET	23	possible 4-hydroxybenzoyl-CoA thioesterase
RPA1812		24467.277, W		1xDEM-9xMET	66	conserved hypothetical protein
RPA1824	17461.560, M			1xDEM-1xMET	32	unknown protein
RPA1825	9003.751, M			6xMET	11	conserved hypothetical protein
RPA1827	9668.191, M			1xDEM-3xMET	19	hypothetical protein
RPA1831	9577.858, S			1xDIS-2xMET	20	conserved hypothetical protein
RPA1839	13292.780, W			6xMET	25	putative dihydroneopterin aldolase
RPA1842	12496.787, M			1xDEM-4xMET	14	conserved unknown protein
RPA1855	22351.107, M			1xDIS-5xMET	74	conserved hypothetical protein
RPA1870			18515.305, M	1xDEM-7xMET	82	possible transcriptional regulator (MarR/EmrR
RPA1872	11267.049, M			7xMET	11	Rhodocoxin
RPA1896		19940.857, M		7xMET	38	homologue of Rhodobacter capsulatus gene
RPA1900		10669.437, M	10697.929, S	6xMET 8xMET	95	homologue of Rhodobacter capsulatus gene
RPA1905		31763.171, M		1xDEM-2xDIS-3xMET	50	homologue of Rhodobacter capsulatus gene
RPA1908	16820.322, M			4xMET	77	hypothetical protein
RPA1909		17137.601, M		1xMET	11	putative transcriptional regulator, MarR family
RPA1915	24984.711, M			1xDEM-9xMET	58	FeuP two-component system, regulatory protein
RPA1928	33099.112, M			2xDIS-6xMET	76	ferredoxin-like protein [2Fe-2S]
RPA1964			19754.608, M	1xDEM-4xMET	14	hypothetical protein
RPA1978		37839.176, S		1xDEM-1xDIS-8xMET	29	molybdenum biosynthetic protein A

Table 7.4: Continued

Gene Number	Aerobic	Anaerobic	Nitrogen Fixing	PTMs	BU Seq. Cov.[e]	Protein Annotation
RPA1982	20026.946, S		19882.199, W	8xMET 1xDEM-7xMET	15	conserved unknown protein
RPA1985	12722.134, M			1xDEM-3xMET	12	probable diacylglycerol kinase
RPA1992	15165.192, M			2xMET	88	possible NtrR protein
RPA1993	9832.024, M			1xDEM-3xMET	14	possible virulence-associated protein
RPA1996			10698.920, M	2xDIS-6xMET	16	hypothetical protein
RPA2004	22567.353, W			1xDIS-8xMET	95	conserved hypothetical protein
RPA2006	26621.810, S			1xDEM-8xMET	74	putative phosphatidylserine decarboxylase
RPA2012	11188.935, S	11118.798179, M		1xDEM-5xMET 1xDEM	36	conserved unknown protein
RPA2028	21793.511, S			1xDEM-4xMET	73	conserved hypothetical protein
RPA2032	19942.040, M				89	acetolactate synthase (small subunit)
RPA2036	26939.380, S			1xDIS-2xMET	41	possible transcriptional regulator (GntR family)
RPA2040	22362.657, M			4xMET	74	possible choline ABC transporter ATP-binding
RPA2044	27086.658, M			1xDIS-7xMET	55	conserved unknown protein
RPA2045		36322.528, S		1xDEM-3xDIS-1xMET	51	biotin synthetase
RPA2057	12463.311, W			1xDEM-4xMET	19	hypothetical protein
RPA2066			35090.978, M	1xDEM-4xMET	60	putative nosX
RPA2068	10674.355, M			1xMET	14	conserved unknown protein
RPA2082	25774.205, W			1xDEM-6xMET	22	putative uroporphyrin III methylase
RPA2084	27462.806, M			1xDEM	13	precorrin 3 or 4 methylase
RPA2085	12941.148, W			2xMET	14	cobalamin biosynthesis protein G; CbiG
RPA2125			12587.776, M		24	conserved unknown protein
RPA2136	11243.957, M		11115.419, W	1xDIS-1xMET 1xDEM-1xMET	18	possible cytochrome C precursor
RPA2145	28029.307, S			1xDIS-6xMET	54	putative enoyl-CoA hydratase/isomerase
RPA2158	11824.846, S			1xDEM-2xDIS-3xMET	11	hypothetical protein
RPA2159	11346.894, W			1xDIS-3xMET	13	hypothetical protein

Table 7.4: Continued

Gene Number	Aerobic	Anaerobic	Nitrogen Fixing	PTMs	BU Seq. Cov.[e]	Protein Annotation
RPA2165	11164.094, S	11163.802, S	11164.046, S		27	chaperonin GroES2, cpn10
RPA2179			20109.335, M		87	xanthine-guanine phosphoribosyltransferase
RPA2188			11659.550, M	1xDIS-6xMET	11	hypothetical protein
RPA2196			25015.256, M		89	conserved hypothetical protein
RPA2197		25380.431, M		1xDEM-9xMET	47	cell division protein FtsJ
RPA2205	13441.281,M				11	hypothetical protein
RPA2239	22823.257, M			1xDEM-3xMET	10	putative partition protein
RPA2241		17775.413, M		1xDEM-3xMET	11	conserved hypothetical protein
RPA2243	9127.186, M			5xMET	16	putative transcriptional regulator
RPA2264	17948.706, S			6xMET	85	conserved hypothetical protein
RPA2265		35913.907, W		1xDEM-1xDIS-2xMET	33	conserved hypothetical protein
RPA2274	13462.515, W			9xMET	11	hypothetical protein
RPA2283	10475.250, S			1xDEM-4xMET	14	putative proteic killer suppression protein
		10517.445, W		1xDEM-7xMET		
RPA2313			19275.381, M	5xMET	11	unknown protein
RPA2314			15152.212, M	1xDEM-1xDIS-6xMET	13	cytochrome c556
RPA2334		11959.120, M		1xDEM		unknown protein
			11959.120, M	1xDEM		
RPA2335		11442.593, S		NATIVE, 1-4xMET	76	unknown protein
			11442.593, S	NATIVE, 1-4xMET		
RPA2336		10908.894,S		NATIVE, 1xMET	47	unknown protein
			10921.606, M	NATIVE, 1xMET		
RPA2338		17752.529		1xDEM	52	unknown protein
RPA2359	30819.670, W			4xMET	52	putative periplasmic protein
RPA2368	8395.593, S			2xMET	17	possible transcriptional regulatory protein
RPA2401	12335.292, S				19	conserved unknown protein
RPA2407	12113.494, W			1xDEM-3xMET	12	hypothetical protein
RPA2409		24359.957, S		1xDIS-6xMET	32	possible AmiR antitermination protein
RPA2421	15886.364, M			4xMET	16	NADH:ubiquinone oxidoreductase 17.2 k

Table 7.4: Continued

Gene Number	Aerobic	Anaerobic	Nitrogen Fixing	PTMs	BU Seq. Cov.[e]	Protein Annotation
RPA2433	16394.228, W			4xMET	12	possible two-component response regulator
RPA2437	15964.473, M			1xDEM-1xDIS-6xMET	12	3-dehydroquinate dehydratase type 2
RPA2442	27754.376, M			1xDIS-5xMET	78	putative outer membrane protein
RPA2443		25651.912, W		1xDEM-3xMET	24	probable antioxidant protein
RPA2446		43113.781, S		4xMET	33	putative aminotransferase
RPA2453		35496.288, S		1xDEM-9xMET	37	translation peptide releasing factor RF-2
RPA2456	23830.531, S				31	possible bacterioferritin co-migratory protein
RPA2465	27433.140, S			2xMET	56	sufC, related to ABC transporter ATP-binding
RPA2470	13365.283, M				13	Protein of unknown function, HesB/YadR/YfhF
RPA2492		22366.559, M		1xDEM-1xDIS-2xMET	79	Conserved hypothetical protein
RPA2513	20843.685, S			9xMET	41	elongation factor P
			20843.911, M	9xMET		
RPA2520	20568.489, M			2xMET	32	hypothetical protein
RPA2521	7939.293, S			1xDEM-8xMET	21	hypothetical protein
RPA2522	9477.775, S			3xMET	22	hypothetical protein
RPA2523	17028.379, S			1xDEM-1xDIS	81	putative lactoylglutathione lyase
			17028.555, S	1xDEM-1xDIS		
RPA2528	19053.121, S			2xDIS-1xMET	71	hypothetical protein
RPA2531	10302.373, W			7xMET	10	hypothetical protein
RPA2533	15410.303, M			1xDEM-6xMET	11	unknown protein
RPA2540	31329.289, S			1xDEM-3xDIS-4xMET	43	3-hydroxy-3-methylglutaryl-CoA lyase
RPA2546	15869.205, M			3xMET	15	FKBP-type peptidyl-prolyl cis-trans isomerase
			15966.401, W	10xMET		
RPA2549			16217.439, M	1xDIS-6xMET	10	conserved hypothetical protein
RPA2552	11779.037, S			8xMET	15	unknown protein
RPA2556		30215.805, M		1xDEM-4xMET	36	PA-phosphatase related phosphoesterase
RPA2589	21640.620, M			1xDIS-2xMET	78	possible competence-damaged protein
RPA2601	17688.403, M			1xMET	83	phosphopantetheine adenylyltransferase
		17771.338, S		7xMET		
RPA2603	16010.780, M			2xMET	88	conserved hypothetical protein
RPA2604	16795.230, M			1xDEM	11	peptidyl prolyl cis-trans isomerase

Table 7.4: Continued

Gene Number	Aerobic	Anaerobic	Nitrogen Fixing	PTMs	BU Seq. Cov.[e]	Protein Annotation
RPA2639	17915.266, M			2xMET	13	probable L-2-amino-thiazoline-4-carboxylic acid
RPA2640			25979.119, S	5xMET	71	Isochorismatase hydrolase family
RPA2648	27758.256, M			1xDEM-5xMET	39	unknown protein
RPA2649	12888.806, M			1xDEM-7xMET	15	conserved unknown protein
RPA2652		7178.843, W				unknown protein
RPA2667	26252.611, W			1xDEM-1xDIS-7xMET	81	conserved unknown protein
RPA2687			17032.969, S	2xMET	82	large-conductance mechanosensitive channel
RPA2688	18258.432, M	18114.023, M		8xMET 1xDEM-7xMET	83	small protein B
RPA2690		24467.277, W		1xDEM-3xDIS-5xMET	40	possible uracil-DNA glycosylase
RPA2692			14368.256, M		20	RNA polymerase omega subunit
RPA2695	15655.277, M		15813.404, W	1xDEM 2xMET	15	acyl carrier protein synthase
RPA2702	18944.166, M			8xMET	94	DUF24, predicted transcriptional regulator,
RPA2715	19653.464, M				16	possible transcriptional regulator, MarR family
RPA2717	7852.327, M			1xDEM-6xMET	29	conserved hypothetical protein
RPA2718	21067.252, M			1xDEM-5xMET	14	hypothetical protein
RPA2721	9921.998, M			2xDIS-3xMET	22	hypothetical protein
RPA2728		17040.698, W	17026.731, M	1xDEM-6xMET 1xDEM-5xMET	56	riboflavin synthase, beta chain
RPA2729	19691.488, M			5xMET	86	putative N-utilization substance protein B
RPA2732	8397.260, W		8397.280, M	1xDEM 1xDEM	48	conserved hypothetical protein
RPA2734	32102.932, S			4xMET	31	possible epoxide hydrolase-related protein
RPA2742	12862.989, M		12904.490, M	2xMET 5xMET	22	integration host factor alpha subunit
RPA2744		7862.194, S		1xDEM-2xMET	95	hypothetical protein
RPA2748	34871.554, W			3xDIS-3xMET	53	possible short-chain dehydrogenase
RPA2755			18260.081, M	1xMET	13	possible DNA-binding stress protein
RPA2766	15132.167, W			1xDEM-4xMET	85	Phenylacetic acid degradation-related
RPA2783	21797.403, M			1xDIS-8xMET	11	hypothetical protein

Table 7.4: Continued

Gene Number	Aerobic	Anaerobic	Nitrogen Fixing	PTMs	BU Seq. Cov.[e]	Protein Annotation
RPA2795			28838.153, M	1xDIS-7xMET	10	Protein of unknown function UPF0001
RPA2801	15572.265, M			1xDIS-1xMET	33	Collagen triple helix repeat
			15571.851, M	1xDIS-1xMET		
RPA2814	17696.443, S			1xDEM	46	single-strand DNA-binding protein
			17696.403, S	1xDEM		
RPA2823	10207.451, M				45	conserved hypothetical protein
RPA2848	17799.324, M			1xDEM-5xMET	13	possible sec-independent protein secretion
RPA2852		35915.053, W		1xDEM-5xMET	32	putative sugar hydrolase
RPA2856	11421.239, W			1xDEM-1xDIS-5xMET	83	Protein of unknown function, HesB/YadR/YfhF
RPA2868	11896.347, W			2xMET	14	Septum formation initiator
RPA2869			20102.175, S	1xDEM-1xDIS-6xMET	92	possible flavin-dependent oxidoreductase
RPA2892	18214.481, M			3xMET	57	molybdenum cofactor biosynthesis protein C
RPA2896	8985.992, W			1xDEM-1xDIS-8xMET	13	hypothetical protein
		8986.770, W		1xDEM-1xDIS-8xMET		
			8986.950, M	1xDEM-1xDIS-8xMET		
RPA2899	18139.023, M			3xMET	99	conserved hypothetical protein
		18139.454, W		3xMET		
RPA2919	20660.667, M			1xDEM-2xMET	22	ribosome releasing factor
		20660.757, M		1xDEM-2xMET		
RPA2932			14258.256, S	2xDIS-1xMET	11	hypothetical protein
RPA2933	9923.460, M			1xDEM	22	conserved hypothetical protein
			9923.086, M	1xDEM		
RPA2934	14212.615, W			1xDEM-1xMET	13	conserved unknown protein
RPA2940		11024.113, W		1xDEM-8xMET	12	NADH-ubiquinone dehydrogenase chain K
RPA2942	18761.236, M			8xMET	86	NADH-ubiquinone dehydrogenase chain I
			18685.995,W	3xDIS-3xMET		
RPA2953	11116.797, S			1xDEM-2xMET	33	possible DNA-binding protein hu-alpha (NS2)
			11145.405, S	1xDEM-4xMET		
RPA2966		12335.167, M	12335.891, M		98	nitrogen regulatory protein P-II
RPA2973			20173.446, M	1xDEM	12	hypothetical protein
RPA2982	10002.695, M			4xMET	13	possible insertion element ISR1 hypothetical 10

Table 7.4: Continued

Gene Number	Aerobic	Anaerobic	Nitrogen Fixing	PTMs	BU Seq. Cov.[e]	Protein Annotation
RPA2985			24851.053, M	1xDEM-5xMET	93	conserved unknownl protein
RPA2998	32018.782, W			7xMET	49	putative dihydrodipicolinate synthase
RPA3005	10219.565, M			4xMET	19	hypothetical protein
RPA3021	21125.201, M			3xMET	24	transcriptional regulator
		21139.221, M		4xMET		
RPA3024			14141.632, W	1xDEM-1xDIS-6xMET	17	unknown protein
RPA3034		9212.003, M		1xDEM-2xMET		unknown protein
			9212.352, M	1xDEM-2xMET		
RPA3035	15363.893, M			5xMET	87	hypothetical protein
			15307.924, M	1xMET		
RPA3037			21369.003, S	7xMET	63	conserved unknown protein
RPA3053	7525.886, S	7526.470, S			65	cold shock protein
RPA3056	15289.620, W			5xMET	33	nucleoside-diphosphate-kinase
		15289.332, M		5xMET		
			15289.489,M	5xMET		
RPA3073			8567.662, S	1xDEM-9xMET	16	constitutive acyl carrier protein
RPA3074		25406.064, M		4xMET	57	3-oxoacyl-acyl carrier protein reductase fabG
RPA3077	17932.594, M		17932.615, M		18	possible 30S ribosomal protein S6
RPA3078	9046.970, S			1xDEM-6xMET	13	30S ribosomal protein S18
RPA3080			21178.830, S		31	putative 50S ribosomal protein L9, cultivar
RPA3086	9397.898, S			3xMET	17	hypothetical protein
RPA3101	17232.381, W			1xDEM-7xMET	88	conserved unknown protein
RPA3103	7088.891, W			8xMET	13	hypothetical protein
		7088.003, W		8xMET		
RPA3109			10697.310, M	9xMET	64	conserved hypothetical protein
RPA3113			17140.300, M	4xMET	10	conserved hypothetical protein
RPA3123	11112.651, W			1xDEM-8xMET	83	hypothetical protein
			11056.853, S	1xDEM-4xMET		
RPA3126	15994.192, M			5xMET	10	conserved hypothetical protein
RPA3129	6249.313, S			1xDEM-1xMET	30	50S ribosomal protein L33
		6249.333, S		1xDEM-1xMET		
			6249.462, S	1xDEM-1xMET		

Table 7.4: Continued

Gene Number	Aerobic	Anaerobic	Nitrogen Fixing	PTMs	BU Seq. Cov.[e]	Protein Annotation
RPA3130	14169.554, S			1xDEM-3xMET	79	Helix-turn-helix motif
RPA3134	10997.690, S			9xMET	16	conserved unknown protein
		10998.578, M		9xMET		
			10997.732, S	9xMET		
RPA3148			14709.646, W	1xDEM-7xMET	69	DUF174
RPA3152			25980.423, S	1xDIS-3xMET	57	hypothetical protein
RPA3162	13284.110, M			7xMET	16	possible helix-turn-helix
		13312.971, W		9xMET		
RPA3168	8804.559, S			1xDEM-7xMET	21	possible flgaellar switch protein FliN
RPA3180	11166.904, S			9xMET	11	hypothetical protein
		11139.519, S		7xMET		
RPA3212	15955.683, M			7xMET	14	unknown protein
RPA3213		8030.265, S		1xDEM-6xMET	18	hypothetical protein
RPA3215	25048.560, S			1xDEM-5xMET	80	putative nitroreductase
RPA3223			35934.629, S	1xDEM-7xMET	68	putative alginate lyase
RPA3225	15717.489, S			3xMET	21	50S ribosomal protein L17
			15717.933, S	3xMET		
RPA3227	13761.002, M			1xDEM-1xMET	22	30S ribosomal protein S11
		13761.158, M		1xDEM-1xMET		
			13761.287, M	1xDEM-1xMET		
RPA3228	14357.764, M			1xDEM-3xMET	79	30S ribosomal protein S13
		14357.709, M		1xDEM-3xMET		
			14384.163, M	1xDEM-5xMET		
RPA3231	16837.254, S	16836.541, S	16836.630, S		87	50S ribosomal protein L15
RPA3232	7093.155, S			1xDEM	20	ribosomal protein L30
		7092.929, S		1xDEM		
			7092.861, S	1xDEM		
RPA3234			12905.679, S	1xDEM	26	50S ribosomal protein L18
RPA3235	19273.207, S			1xDEM	41	50S ribosomal protein L6
		19273.083, S		1xDEM		
			19273.221, S	1xDEM		
RPA3236	14576.761, S			1xDEM-7xMET	27	30S ribosomal protein S8
			14576.574, S	1xDEM-7xMET		
RPA3238	21122.037, M			6xMET	12	50S ribosomal protein L5
RPA3239	11012.697, S			1xDEM-1xMET	21	50S ribosomal protein L24
		11012.444, S		1xDEM-1xMET		
			11012.821, S	1xDEM-1xMET		
RPA3240	13489.094, M	13488.674, M	13489.187, M		22	50S ribosomal protein L14
RPA3242	7849.937, S			1xDEM	24	50S ribosomal protein L29
		7849.874, S		1xDEM		
			7849.864, S	1xDEM		

Table 7.4: Continued

Gene Number	Aerobic	Anaerobic	Nitrogen Fixing	PTMs	BU Seq. Cov.[e]	Protein Annotation
RPA3243	15296.048, S			1xMET	23	50S ribosomal protein L16
			15296.084, S	1xMET		
RPA3244	26178.851, W			1xDEM-1xDIS-3xMET	60	30S ribosomal protein S3
RPA3246	10088.469, S			1xDEM	27	30S ribosomal protein S19
		10088.391, S		1xDEM		
			10088.778, S	1xDEM		
RPA3248	10908.765, S	10908.483, S	10908.697, S		31	50S ribosomal protein L23
RPA3251	11668.941, S	11668.783, S	11668.568, S		40	30S ribosomal protein S10
RPA3254			17556.347, M	1xDEM	27	30S ribosomal protein S7
RPA3261	15004.785, W			1xDIS-1xMET	91	transcriptional regulator
			14902.234, M	1xDEM-1xDIS-3xMET		
RPA3269	12754.760, S			1xDEM-3xMET	47	50S ribosomal protein L7/L12
		12754.702, S		1xDEM-3xMET		
			12754.772, S	1xDEM-3xMET		
RPA3270	19054.936, M			1xDEM-1xMET	28	50S ribosomal protein L10
RPA3272	23878.783, S			1xDEM	12	50S ribosomal protein L1
RPA3274	20026.946, S			9xMET	39	transcription antitermination protein
		20026.860, S		9xMET		
RPA3275			9402.074, M	5xMET	84	preprotein translocase, SecE subunit
RPA3290	23849.470, S			1xDEM-1xDIS-3xMET	32	possible transcriptional regulator, TetR family
RPA3300	27834.970, S			2xMET	73	possible transcriptional regulator, TetR family
RPA3319	15533.356, W			7xMET	58	hypothetical protein
		15560.493, S		9xMET		
			15532.882, S	7xMET		
RPA3327	10062.561, W			1xDEM-4xMET	16	hypothetical protein
RPA3328			15602.023, W	5xMET	13	conserved hypothetical protein
RPA3373	10271.970, M			7xMET	27	hypothetical protein
RPA3390	16873.403, M			1xDIS-8xMET-	16	phosphoribosyl c-AMP cyclohydrolase
		16789.906, M		1xDIS-2xMET		
			16672.650, M	1xDEM-1xDIS-3xMET		
RPA3394		17200.646, S		1xDEM-1xDIS-3xMET	93	DUF37
RPA3397		16658.898, M		6xMET	11	hypothetical protein
RPA3402	15218.140, W			1xDEM-2xDIS-5xMET	15	conserved hypothetical protein

Table 7.4: Continued

Gene Number	Aerobic	Anaerobic	Nitrogen Fixing	PTMs	BU Seq. Cov.[e]	Protein Annotation
RPA3434	22791.311, M			1xDIS-6xMET	47	Universal stress protein (Usp)
RPA3436	18900.382, M			1xDEM-2xMET	13	GCN5-related N-acetyltransferase
RPA3446	29567.345, M			1xDIS-2xMET	51	3-hydroxyisobutyrate dehydrogenase
			29595.953, M	1xDIS-4xMET		
RPA3457			7429.588, S	1xDEM-5xMET	20	Biotin/lipoyl attachment:Biotin-requiring
RPA3476	28083.780, S			1xDEM-2xMET	84	possible energy transducer TonB
RPA3481	9215.385, S			7xMET	23	hypothetical protein
RPA3501	10149.437, W			1xDEM-7xMET	67	conserved unknown protein
		10163.493, W		1xDEM-8xMET		
RPA3518	11231.845, W			4xMET	95	Excinuclease ABC, C subunit, N-terminal
RPA3524		36319.974, S		1xDEM-6xMET	43	putative cell division protein FtsQ
RPA3537	14096.695, M			1xMET	11	conserved hypothetical protein
RPA3555	18816.879, S			1xDEM-2xMET	92	arsenate reductase
RPA3561		13124.930, S		1xDIS-8xMET	14	possible arsenate reduction regulatory protein
RPA3574	6883.790, M			1xDEM-4xMET	32	putative thiamin biosynthesis ThiG
		6897.073, M		1xDEM-5xMET		
RPA3579	10002.695, W			4xMET	13	possible insertion element ISR1 hypothetical 10
		10030.489, W		6xMET		
RPA3583	12099.680, W		12127.793, M	1xDEM-5xMET	22	conserved hypothetical protein
RPA3589	RPA3589			1xDEM-7xMET		conserved hypothetical protein
RPA3602			9716.998, W	1xDIS-4xMET	25	unknown protein
RPA3606	10500.755, W			1xDIS-1xMET	13	hypothetical protein
RPA3626			9293.434, M	1xDEM	27	conserved unknown protein
RPA3652			13698.826, S	1xDEM-2xMET	69	conserved hypothetical protein
RPA3653			9296.464, W	4xMET	26	Protein of unknown function UPF0033
RPA3662	11279.887, W			6xMET	14	urease beta subunit
RPA3663	10999.640, S			1xDEM-3xMET	20	urease gamma subunit
		10998.597, S		1xDEM-3xMET		
RPA3671	10459.920, S			1xDEM-1xMET	59	translation initiation factor if-1 (infA)
		10460.115, S		1xDEM-1xMET		
			10459.441, S	1xDEM-1xMET		

Table 7.4: Continued

Gene Number	Aerobic	Anaerobic	Nitrogen Fixing	PTMs	BU Seq. Cov.[e]	Protein Annotation
RPA3672	8970.250, S	8969.587, M	9012.829, M	1xDEM-2xMET 1xDEM-2xMET 1xDEM-5xMET	47	cold shock protein
RPA3673	22674.886, M			8xMET	82	conserved unknown protein
RPA3676			18594.783, M	1xDEM-1xDIS-1xMET	35	putative type IV prepilin peptidase, cpaA
RPA3719	28867.609, S			1xDEM-6xMET	64	putative high-affinity branched-chain amino acid
RPA3721		30808.822, W		1xDEM-7xMET	45	possible ABC transporter, permease protein
RPA3726	18517.400, M			6xMET	90	conserved unknown protein
RPA3745	17028.339, S			5xMET	90	unknown protein
RPA3759			14695.257, M	1xDEM-7xMET	31	putative 5-carboxymethyl-2-hydroxymuconate
RPA3770	17958.517, M			1xDEM-3xMET	13	conserved unknown protein
RPA3786	9335.393, S	9335.064, S			57	unknown protein
RPA3790	28964.617, M			1xDIS-9xMET	56	putative efflux protein
RPA3794	11869.466, S			4xDIS-4xMET	17	conserved hypothetical protein
RPA3798	24649.796, M				34	conserved unknown protein
RPA3799	11347.234, S	11347.449, M		1xDIS-4xMET 1xDIS-4xMET	16	DUF182
RPA3803	17136.725, S			1xDEM-1xDIS-8xMET	50	carbon-monoxide dehydrogenase small subunit
RPA3804	16026.165, S		16083.263, M	PTM:1xDIS 1xDIS-4xMET	35	conserved unknown protein
RPA3820	8418.840, M		8419.307, M		29	Protein of unknown function UPF0062
RPA3822	12010.648, W			5xMET	13	conserved unknown protein
RPA3824	12572.842, M			1xDEM	13	conserved hypothetical protein
RPA3826			13316.091, M	9xMET	12	conserved hypothetical protein
RPA3827			11761.786, M	3xMET	11	conserved hypothetical protein
RPA3828			7145.348, M	1xDEM-2xMET	25	Helix-turn-helix motif
RPA3837	10002.695, W	10030.489, W		4xMET 6xMET	13	possible insertion element ISR1
RPA3852	20592.106, M			1xDEM-8xMET	86	hypothetical protein

Table 7.4: Continued

Gene Number	Aerobic	Anaerobic	Nitrogen Fixing	PTMs	BU Seq. Cov.[e]	Protein Annotation
RPA3864	9463.912, S	9463.471, S			19	unknown protein
RPA3865		19743.894, S		2xMET	68	Thioesterase superfamily
RPA3871			26980.102, S	1xDIS-8xMET	29	Nuclear protein SET
RPA3875	9961.592, M			1xDEM	76	conserved unknown protein
		9961.076, M		1xDEM		
			9961.140, M	1xDEM		
RPA3878	17694.943, S			1xDEM-5xMET	74	conserved unknown protein
RPA3886			10475.058, W	5xMET	12	Flagellar hook-basal body complex protein FliE
RPA3887	15471.115, M			5xMET	78	flagellar basal-body rod protein flgC
RPA3895	28030.834, M			6xMET	59	conserved hypothetical protein
RPA3896			15002.200, W	1xDEM-3xMET	10	hypothetical protein
RPA3898	18030.359, M			3xMET	11	Flagellar basal body-associated protein FliL
RPA3907	16930.802, M			9xMET	14	DnaK suppressor protein DksA
RPA3908	14169.336, S			8xMET	88	conserved unknown protein
RPA3910	12130.745, S			1xMET	20	conserved hypothetical protein
		12130.782, M		1xMET		
			12186.408, W	5xMET		
RPA3913	13458.449, S			1xDEM-7xMET	74	conserved hypothetical protein
RPA3914	14420.772, M			1xDEM-1xMET	12	putative flbT protein
RPA3923		34837.447, W		1xDEM-2xDIS-3xMET	56	putative acetoin dehydrogenase (TPP-dependent)
RPA3924	24964.302, S			1xDEM-5xMET	87	conserved hypothetical protein
RPA3939	19601.710, S			1xDIS-7xMET	97	conserved unknown protein
RPA3956	11442.106, S			1xDEM-6xMET	30	ferredoxin
			11442.720, S	1xDEM-6xMET		
RPA3957		12561.372, M		2xMET	84	Hpt domain
RPA3970	23794.943, S			1xDEM-1xMET	69	putative
RPA3988	19992.661, S			1xDEM-3xMET	95	putative phosphatase
RPA4005	10845.866, S			1xMET	12	possible ribosomal protein S21
RPA4006		17769.328, S		1xDEM-5xMET	12	hypothetical protein
RPA4010		14013.813, M		3xMET	13	putative response regulator
RPA4030	13980.157, W			4xMET	13	hypothetical protein

Table 7.4: Continued

Gene Number	Aerobic	Anaerobic	Nitrogen Fixing	PTMs	BU Seq. Cov.[e]	Protein Annotation
RPA4047		25162.285, W		2xDIS-6xMET	59	Haloacid dehalogenase-like hydrolase
RPA4050	29896.503, S			1xDEM-3xMET	25	unknown protein
			29896.082, S	1xDEM-3xMET		
RPA4067	11526.438, S			1xDEM-3xMET	13	hypothetical protein
RPA4070	21755.278, M			1xDEM-3xMET	50	possible peptide methionine sulfoxide reductase
RPA4072	17029.108, S			1xDEM	61	transcriptional elongation factor greA
		17029.156, S		1xDEM		
			17029.160, S	1xDEM		
RPA4074	17903.487, S			1xDIS-3xMET	69	putative leucine regulon transcriptional
		17987.102, M		1xDIS-9xMET		
RPA4076	34867.931, W			4xMET	49	putative transcriptional regulator, lysR family
RPA4077	23944.810, M			1xDEM-1xDIS-7xMET	76	ATPase, ParA type
RPA4093			9159.405, M	1xDEM-3xMET	14	hypothetical protein
RPA4102	17031.513, M			1xDEM-1xMET	12	putative transcriptional regulator
		17086.813, S		1xDEM-5xMET		
RPA4104	22887.971, M			7xMET	66	hypothetical protein
RPA4109		11913.973, S		7xMET	67	conserved hypothetical protein
			11913.701, M	7xMET		
RPA4122	9250.851, M			5xMET	12	Conjugal transfer protein TrbD
RPA4129	12832.671, S			9xMET	30	putative transcriptional regulator
		12832.592, S		9xMET		
			12832.545, S	9xMET		
RPA4135		18385.213, S		5xMET	14	GCN5-related N-acetyltransferase
RPA4137		13892.225, S		1xDEM	20	conserved unknown protein
RPA4138	13905.092, S			1xDEM-1xMET	28	conserved unknown protein
RPA4151		31499.295, W		2xDIS-2xMET	34	possible transcriptional regulator of NADH
RPA4171	12340.401, W			3xMET	19	conserved unknown protein
		12341.050, M		3xMET		
			12326.933, W	2xMET		
RPA4176	10077.843, M			1xMET	12	ribosomal protein S21
RPA4179		13333.444, S		1xDEM-5xMET	57	conserved unknown protein
RPA4191	11212.504, S			1xDEM-7xMET	28	conserved unknown protein
RPA4206	27433.140, S			1xDEM-2xMET	77	D-beta-hydroxybutyrate dehydrogenase

Table 7.4: Continued

Gene Number	Aerobic	Anaerobic	Nitrogen Fixing	PTMs	BU Seq. Cov.[e]	Protein Annotation
RPA4210			9267.744, W	4xMET	20	hypothetical protein
RPA4214	10761.067, M			4xMET	14	conserved hypothetical protein
RPA4217			7991.877, S		35	conserved unknown protein
RPA4224	10459.824, S			6xMET	35	unknown protein
RPA4227	22384.074, S			1xDIS-4xMET	67	urease accessory protein UreG
RPA4228	12550.962, W			1xDEM-4xMET	15	hypothetical protein
RPA4230	10046.148, S		10073.371, M	1xDEM-1xMET 1xDEM-3xMET	13	conserved unknown protein
RPA4241		15050.666, S		1xDEM-6xMET	70	CBS domain
RPA4257	19545.573, M			2xDIS-4xMET	46	NADH-ubiquinone dehydrogenase chain I
RPA4272			25979.575, S	1xDEM	15	conserved unknown protein
RPA4277	23149.083, S			1xDEM-3xMET	48	conserved hypothetical protein
RPA4282	18029.260, W			6xMET	70	possible activator of photopigment and puc
RPA4297	30057.486, W			1xDEM-1xDIS-5xMET	60	putative aldose reductase
RPA4298	22791.264, S			2xMET	68	ATP/GTP-binding site motif A (P-loop)
RPA4305	12110.554, M			1xDEM-5xMET	10	hypothetical protein
RPA4319	5806.499, W			4xMET	25	hypothetical protein
RPA4330	17933.468, S			2xMET	13	conserved unknown protein
RPA4331		44426.051, S			46	aspartate aminotransferase A
RPA4344	8492.921, M			1xDEM-6xMET	22	hypothetical protein
RPA4348		16278.313, M		1xDEM-5xMET	70	conserved hypothetical protein
RPA4349	18866.714, S		18851.297, M	1xDEM-1xMET 1xDEM-1xDIS	20	conserved unknown protein
RPA4357	10095.787, M			1xDEM-1xMET	11	conserved unknown protein
RPA4365			15801.209, M	1xDEM-1xDIS	13	GCN5-related N-acetyltransferase
RPA4372	15491.980, M			4xMET	10	Class I peptide chain release factor domain
RPA4381	16826.447, M			1xDIS-6xMET	59	conserved unknown protein
RPA4383	18847.915, M	18804.927, M		1xDEM-6xMET 1xDEM-3xMET		conserved unknown protein
RPA4393			7991.020, M	4xMET	20	unknown protein

Table 7.4: Continued

Gene Number	Aerobic	Anaerobic	Nitrogen Fixing	PTMs	BU Seq. Cov.[e]	Protein Annotation
RPA4418			7110.894, W	1xDEM-9xMET	22	conserved unknown protein
RPA4457	13639.667, M			1xDEM	20	putative sulfide dehydrogenase
RPA4458	17466.002, W			5xMET	51	hypothetical protein
RPA4466	12125.648, M			1xDEM	15	putative sulfur oxidation protein soxZ
		12125.666, M		1xDEM		
RPA4467	16177.138, M			1xDEM-2xMET	98	putative sulfur oxidation protein soxY
RPA4470	16387.876, W			1xDEM-3xMET	12	DUF336
RPA4473	11040.704, S			4xMET	27	conserved hypothetical protein
RPA4474	14375.557, S			7xMET	12	possible transcriptional activator
			14316.966, W	1xDIS-3xMET		
RPA4478	10792.071, S			1xMET	13	conserved hypothetical protein
		10805.846, M		2xMET		
RPA4483		43079.763, S		1xDEM-1xDIS-7xMET	29	possible signal transducer
RPA4500	7526.018, S			1xDEM-1xDIS-5xMET	32	hypothetical protein
RPA4501		7962.293, M		2xMET	20	phnA-like protein
RPA4503	9513.824, M			1xDIS-7xMET	14	conserved hypothetical protein
RPA4505		26141.064, M		1xMET	56	TPR repeat
RPA4518	11898.625, M			2xMET	20	hypothetical protein
RPA4529	17419.338, M			1xDEM-5xMET	10	putative arsenate reductase
RPA4541	19273.064, S			1xDEM-2xMET	73	DNA invertase gene rlgA
RPA4542			11244.156, M	1xDEM-4xMET	18	unknown protein
RPA4543	15054.561, W			8xMET	12	conserved hypothetical protein
RPA4544	21753.512, M			1xDEM-4xMET	61	conserved unknown protein
RPA4548			12793.518, W	6xMET	52	hypothetical protein
RPA4573		16639.503, M		1xDIS-6xMET	12	conserved unknown protein
RPA4574			9800.818, S	1xDEM-5xMET	26	hypothetical protein
RPA4600	16584.872, M			1xDEM-2xDIS-5xMET	15	conserved unknown protein
RPA4602			10657.886, S	1xDEM-2xDIS-4xMET	11	ferredoxin like protein, fixX
RPA4605			30483.750, S	1xDEM-1xDIS-8xMET	46	electron trnasfer flavoprotein beta chain fixA
RPA4610	10830.525, S			1xDEM-6xMET	18	Protein of unknown function, HesB/YadR/YfhF

Table 7.4: Continued

Gene Number	Aerobic	Anaerobic	Nitrogen Fixing	PTMs	BU Seq. Cov.[e]	Protein Annotation
RPA4612			11212.001, S		16	ferredoxin 2[4Fe-4S] III, fdxB
RPA4615			14719.802, S	7xMET	25	nitrogenase molybdenum-iron protein nifX
RPA4634		30215.805, M		1xDEM-6xMET	28	hypothetical protein
RPA4666	16786.288, S	16786.506, M		1xDIS-4xMET 1xDIS-4xMET	51	carbon-monoxide dehydrogenase small subunit
RPA4676	18896.116, S	18987.873, S		1xDEM-1xDIS-7xMET 4xMET	55	putative transcriptional regulator
RPA4678		22847.204, S		1xDEM-8xMET	13	possible outer membrane protein OprF (AF117972)
RPA4686	28332.691, W			1xDEM-9xMET	68	possible ABC transporter, periplasmic amino
RPA4689		16547.471, M		1xDEM-1xDIS-4xMET	40	conserved hypothetical protein
RPA4701	14599.861, W			1xDEM-4xMET	13	Protocatechuate 4,5-dioxygenase, alpha chain
RPA4703			35090.978, M	2xDIS-6xMET	94	4-carboxy-2-hydroxymuconate-6-semialdehyde
RPA4705	14571.402, M	14530.074, M		1xDEM-7xMET 1xDEM-4xMET	13	phosphoribosyl-AMP cyclohydrolase /
RPA4707	23061.022, M			4xMET	76	conserved hypothetical protein
RPA4724	19977.214, S			1xDEM-3xMET	75	putative uridine 5-monophosphate synthase
RPA4738	13453.223, S		13467.290, M	1xDEM-7xMET 1xDEM-8xMET	18	possible two-component system reponse regulator
RPA4740			14258.256, S	1xDIS-8xMET	11	putative 4-carboxymuconolactone decarboxylase
RPA4744			24159.019, W	7xMET	53	possible thioredoxin-like protein
RPA4748		31898.837, S		5xMET	38	3-hydroxybutyryl-CoA dehydrogenase
RPA4760	18312.327, S			1xDEM	14	unknown protein
RPA4768	28138.586, S	28138.352, M		3xMET 3xMET	39	conserved hypothetical protein
RPA4770	14601.733, M			1xDEM-3xMET	13	DUF525
RPA4774		23446.932, W		1xDIS-5xMET	10	conserved unknown protein
RPA4775		26963.040, W		2xMET	6	phosphate regulatory protein, PhoB

Table 7.4: Continued

Gene Number	Aerobic	Anaerobic	Nitrogen Fixing	PTMs	BU Seq. Cov.[e]	Protein Annotation
RPA4777	30253.206, M			1xDIS-1xMET	13	phosphate ABC transporter ATP-binding protein,
RPA4804	10400.837, M			5xMET	20	conserved hypothetical protein
RPA4818	12907.825, M			3xMET	80	conserved hypothetical protein
RPA4825	13087.366, W			2xMET	11	putative transcriptional regulator, MerR family
RPA4827	8356.579, M		8355.656, M	1xDIS-4xMET 1xDIS-4xMET	20	conserved hypothetical protein
RPA4836	9578.187, S			1xDEM	19	30S ribosomal protein S20

[a]Average molecular weight (Da)
[b]S = Significant signal within the mass spectrum
[c]M = Moderate signal within the mass spectrum
[d]W = Weal signal within the mass spectrum
[e]Percent bottom-up sequence coverage
DEM = N-terminal methionine truncation
ACE = acetylation
MET = Methylation
DIS = Disulfide Bond
URY = Uridylylation

Functional Categories

The use of protein identifications that have only top-down and bottom-up identification limits the results to only the overlapping regions of the two data sets. For example, the top-down analysis will be able to see proteins ≤ 40-60 kDa in size due to the limitation of on line C4 reverse phase chromatography employed and the ability to elute larger proteins off the column. In the bottom-up analysis smaller proteins will be missed, because when the tryptic digest is performed smaller proteins will generate small peptides that are not seen within the mass spectrometer. Therefore, we have a subset of 599 proteins containing both top-down and bottom-up confirmations.

These 599 proteins range in functional categories they belong to. The functional categories for the identified proteins are shown in Table 7.5 (these functional categories are based on the ORNL annotation scheme for bacteria (http://genome.ornl.gov/microbial/). Table 7.5 depicts proteins identified from each category, the total number of proteins predicted in each category from the genome, and the percent of the predicted genome identified from each category. A total of 599 proteins were confidently identified representing 12.44% of the genome predictions. Most of the identified proteins fall into the unknowns and unclassified functional category in Table 7.5. This category contains two sub-groups that includes hypothetical and conserved hypothetical proteins, as well as unknown and conserved unknown. Most of the identified proteins were hypothetical and conserved hypothetical proteins, 141 in total being followed by 100 proteins in the unknown function category.

Table 7.5: Functional categories of identified proteins.

Category	Proteins	Genome Prediction	% Identified
Unknowns and Unclassified	241	1407	17.13
Replication Repair	17	126	13.49
Energy Metabolism	35	306	11.44
Carbon and Carbohydrate metabolism	6	107	5.61
Lipid Metabolism	9	158	5.70
Transcription	54	283	19.08
Translation	56	168	33.33
Cellular Processes	59	524	11.26
Amino Acid Metabolism	17	181	9.39
General Function Prediction	44	420	10.48
Metabolism of Cofactors and Vitamins	16	150	10.67
Transport	17	699	2.43
Signal Transduction	21	231	9.09
Purine and Pyrimidine Metabolism	7	56	12.50
Total	599	4816	12.44

In our classification scheme, protein names are changed from hypothetical and conserved hypothetical to unknown and conserved unknown when they are confidently identified with at least two unique peptides [121]. Another category with numerous identifications includes proteins involved in cellular processes such as chaperones, flagellar proteins, stress proteins, and proteases. This category contained 59 proteins. The *R. palustris* genome contains two separate copies of GroEL (RPA1140 and RPA2164) and GroES (RPA1141 and RPA2165). We identified each of the two predicted GroES proteins encoded by the RPA1141 and RPA2165 genes at high confidence. The two GroEL proteins were not found due to the larger size of these proteins at 57626 Da for GroEL-1 and 57796 Da for GroEL-2. These larger molecular masses will prevent them from being identified in the top-down analysis and therefore excluded from our combined top-down and bottom-up data set.

The categories of transcription and translation make up two of the largest percentages of proteins identified based on genome predictions and were identified in all three growth states. This is to be expected since many of the proteins in these categories are necessary under all metabolic modes. The large number of ribosomal proteins provides most of the identifications in the translation functional category. In a previous study of the purified 70S ribosome from *R. palustris*, we identified 53 of the 54 predicted ribosomal proteins [54]. In the present study, we identified 45 of the 54 predicted ribosomal proteins without prior purification. The missed ribosomal proteins are all small and rich in lysine residues, which suggest that they were digested into peptides too small for confident bottom-up identification. The larger ribosomal proteins were likely missed by the top-down analysis due to the inability to elute them from the reverse phase C4

column. This problem was also encountered in the previous study by Strader et al [54]. The transcription functional category mainly consisted of transcriptional regulators with a total of 24 of these proteins comprising the 54 total identified proteins.

Proteins from the functional categories of, replication and repair, energy metabolism, and purine and pyrimidine metabolism comprise some of the larger percentages of the genome predictions based on the integrated top-down and bottom-up data set. The category of replication and repair was detected with 17 proteins. In a previous study baseline proteomics study performed in our laboratory, replication and repair was found to be the category with the lowest abundance at 17% and 22 identified proteins [121]. Howerver, in this study using the integrated top-down and bottom-up method, the category of replication and repair was one of the higher percentage categories with 17 proteins identified. This total is within the same range as identified in the baseline bottom-up proteomics study. The identification of 17 proteins from this category may be due to the size and ability of these proteins to be eluted from the C4 reverse phase column to be identified by top-down analysis well. In the category of energy metabolism 35 proteins were identified. Most of these identifications include proteins involved in photosynthesis and oxidative phosphorylation. One set of proteins identified of particular interest are the NADH-ubiquinone dehydrogenase complexes. Within this complex two operons (RPA2937-RPA2952 and RPA4252-RPA4264) are each predicted to encode complete NADH-ubiquinone dehydrogenase proteins. It has been theorized, the structure of each operon and the degree of divergence of the individual proteins within these operons have different evolutionary lineages, possibly from lateral transfer instead of duplication and divergence within the genome [13, 121].

A total of 3 proteins were identified from the first operon and 1 protein from the second operon. These proteins were found across all metabolic states indicating expression under all metabolic states. Isoforms of this protein were also identified which will be discussed in greater detail later.

The categories of general function, metabolism of cofactors and vitamins, signal transduction, amino acid metabolism, carbon and carbohydrate metabolism, and lipid metabolism were identified with some of the smallest numbers of proteins predicted by the genome sequence. From the signal transduction category 21 proteins were identified with some of these identifications coming from nitrogen regulation proteins such as the GlnK proteins and GlnB (RPA0272, RPA0274, and RPA2966) as well as the chemotaxis proteins. Proteins from the categories of metabolism of cofactors and vitamins, amino acid metabolism, carbon and carbohydrate metabolism, and lipid metabolism contained proteins expected in metabolism of the individual products from the associated pathways and in most cases represented across all growth states.

The transport category was identified with 17 proteins. This category should be fairly abundant within the proteome. Originally, this was the case in the baseline study of *R. palustris*; the integrated top-down and bottom-up method employed here provides a low percentage of these proteins [121]. This is again due to the large size of the proteins within this category at 30-50 kDa which would prevent the elution during the LC-FT-ICR experiments off the C4 reverse phase column.

Comparison of Growth States

One goal of this study was to identify protein differences between the three growth conditions (aerobic, anaerobic, and nitrogen fixing) employed for *R. palustris*. This

comparison was first done by binary comparisons of related metabolic states as illustrated in Figure 7.1. Proteins identified as showing expression differences between metabolic states were then compared across all metabolic states to determine global trends in protein expression. These differences were based on the presence or absence of the protein in one state as compared to the closest metabolic state. For example, the chemoheterotropic growth state (aerobic) was compared to the photoheterotrophic growth state (anaerobic) as a baseline comparison. Also compared were the photoheterotrophic growth state (anaerobic) and the photoheterotrophic nitrogen fixing growth state. It should be noted that this technique is only useful in determining proteins presence or absence between growth states and generating hypotheses about these proteins for future testing.

Chemoheterotrophic Growth State Compared to the Photoheterotrophic Growth State

The chemoheterotrophic and photoheterotrophic states are the base states for this study, as shown in Figure 7.1. The initial expectation is that the protein profiles of cells grown under these two conditions would be quite different due to the cells obtaining energy from the oxidation of succinate during chemoheterotrophic growth and energy from light during photoheterotrophic growth. Most importantly, chemoheterotrophic cells were grown aerobically whereas photoheterotrophic cells were grown anaerobically. Succinate was the source of carbon for both growth modes. Interestingly, the hallmark *R. palustris* phenotype of photosynthesis, the red coloring of the cell membranes, was observed for every metabolic state, though the red coloring was much more prominent under anaerobic states. This is due to *R. palustris* inability to turn off its photosynthetic machinery completely, no matter what growth condition it is in. Therefore, certain

photosystem proteins would be expected throughout all growth conditions. The photosystem proteins are generally large in size and not amenable to the liquid chromatography separations used in the top-down measurements and therefore, not listed in the total 599 proteins identified. However, certain photosystem proteins, such as RPA1548 which encodes for the H subunit of the photosynthetic reaction center, were identified in the bottom-up analysis across all three growth states. Nonetheless, differences were found at the protein level between these two growth states. The BolA-like protein (RPA0501) and chemotaxis protein CheY4 proteins (R1175) showed strong correlation with the aerobic states with no expression under any of the anaerobic states. In contrast, the anaerobic proteins unknown proteins RPA1495, RPA1620, RPA2333, RPA2334, RPA2335, RPA2336, and RPA2338 all showed strong correlation with the anaerobic states and no expression under the aerobic state. The operon of genes encoding unknown proteins, from RPA2333 to RPA2338, is a unique operon that was previously identified in a baseline proteomics study performed on *R. palustris* [121]. As in the previous study, this entire operon, except RPA2337, was found to show relatively strong expression under anaerobic states but no expression in the aerobic state. In the case of RPA2337 it was not detected, even though it does not have any predicted transmembrane domains which should make it detectable in both the peptide and protein form [121].

Photoheterotrophic Growth State Compared to Nitrogen-Fixing Growth State

In this study the evaluation of protein differences between with nitrogen fixation in the photoheterotrophic state was a logical step in testing our methodology, due to many of the proteins expressed during nitrogen fixation should be present when this process is undertaken by the cell [121, 122]. As expected, a number of proteins expressed

only under nitrogen fixing conditions were identified. Some of the proteins thought to be involved in nitrogen fixation were found only under nitrogen-fixing conditions and not detected to under any of the other growth conditions. These include RPA4209, glutamine synthetase; and certain proteins within the nif regulon (RPA4602-4632) including RPA4605, electron trnasfer flavoprotein beta chain fixA; RPA4612, the ferredoxin 2[4Fe-4S] III, fdxB; as well as RPA4615 nitrogenase molybdenum-iron protein nifX. The protein RPA4209, glutamine synthetase, is involved in nitrogen fixation in concert with the GlnK and GlnB proteins, which are regulated by a unique PTM under nitrogen fixing conditions.

Post Translational Modifications

Of the 599 proteins identified by top-down and bottom-up most of these were identified with some varying degree of PTMs; whether it is an N-terminal methionine truncation, methylation, or acetylation. Nearly all proteins undergo some form of post translational modification [1]. These post translational modifications are important to provide protein heterogeneity; thereby allowing the protein to exists in multiple isoforms. Within this study, the common PTMs of methylation, acetylation, N-terminal methionine truncation, and disulfide bonds were examined. By far the most common PTM identified was N-terminal methionine truncation. Of the 599 proteins identified in this study 267 have a methionine truncation. The truncation of the N-terminal methionine depends on the charge and size of the amino acid side chain occupying the next position from the N-terminal methionine. According to the "N-end rule", residues bearing small uncharged side chains (stabilizing), such as alanine, allow docking of methionine peptidases that cleave the N-terminal methionine. Within the 267 N-terminal truncated proteins

identified, 184 have a stabilizing amino acid according to the "N-end rule". Shown in Table 7.6 are all of the N-terminal methionine truncated proteins identified and the amino. acid occupying the second position. The proteins that do not follow the "N-end rule" would make good candidates for further study. Also the proteins that do not adhere to the rule may be the result of annotation errors and bear further analysis of the gene start site calls.

Phosphorylation is a common PTM, although, most of the phosphorylation in *R. palustris* is performed by a histidine kinase, which provides a very fleeting interaction as well as being acid labile that posses problems during mass spectrometry analysis. Due to these reasons, phosphorylation was not searched for within this study. Other specialized PTMs such as uridylylation were searched for and identified within the top-down and bottom-up data sets.

A number of proteins were identified with PTMs from the anaerobic growth state. Of the 119 proteins identified by the integrated top-down and bottom-up analysis 90 of these proteins from the anaerobic growth were identified with a form of a PTM. The most abundant of the PTMs seen on the 90 proteins are N-terminal methionine truncation, followed by proteins containing combinations of methylations. Of particular interest were the unknown and hypothetical proteins that contain PTMs, due to the possible information about function and location this can provide [Table 7.7]. The unique hypothetical operon (RPA2333-RPA2338) was of interest because it was located in one operon that was previously unknown. It has also been demonstrated that none of the proteins in this operon have been found to have strong similarity to any genes in sequenced microbial genomes to date except RPA2333, which is similar to segments of a

Table 7.6: N-Terminal Methionine Truncations

Gene	Second AA	Function
RPA0008	A	circadian clock protein
RPA0038	A	ribosomal protein L20
RPA0159	A	ribosomal protein L27
RPA0335	A	putative phospholipid N-methyltransferase
RPA0366	A	unknown protein
RPA0526	A	50S ribosomal protein L32
RPA0662	A	ferredoxin
RPA0673	A	transcriptional activator
RPA0927	A	probable transcriptional regulator
RPA0953	A	possible exodeoxyribonuclease small subunit
RPA1088	A	hypothetical protein
RPA1090	A	possible nitrogen regulator
RPA1141	A	chaperonin GroES1, cpn10
RPA1175	A	chemotaxis protein CheY4
RPA1717	A	hypothetical protein
RPA1777	A	DUF35
RPA2012	A	conserved unknown protein
RPA2197	A	cell division protein FtsJ
RPA2334	A	unknown protein
RPA2437	A	3-dehydroquinate dehydratase type 2
RPA2556	A	PA-phosphatase related phosphoesterase
RPA2604	A	peptidyl prolyl cis-trans isomerase
RPA2728	A	riboflavin synthase, beta chain
RPA2768	A	ribosomal protein S9
RPA2814	A	single-strand DNA-binding protein
RPA2852	A	putative sugar hydrolase
RPA2869	A	possible flavin-dependent oxidoreductase
RPA2953	A	possible DNA-binding protein hu-alpha (NS2)
RPA3078	A	30S ribosomal protein S18
RPA3129	A	50S ribosomal protein L33
RPA3227	A	30S ribosomal protein S11
RPA3228	A	30S ribosomal protein S13
RPA3232	A	ribosomal protein L30
RPA3237	A	30S ribosomal protein S14
RPA3238	A	50S ribosomal protein L5
RPA3239	A	50S ribosomal protein L24
RPA3255	A	30S ribosomal protein S12
RPA3270	A	50S ribosomal protein L10
RPA3272	A	50S ribosomal protein L1
RPA3273	A	50S ribosomal protein L11
RPA3436	A	GCN5-related N-acetyltransferase
RPA3457	A	Biotin/lipoyl attachment:Biotin-requiring
RPA3583	A	conserved hypothetical protein
RPA3671	A	translation initiation factor if-1 (infA)
RPA3803	A	carbon-monoxide dehydrogenase small subunit
RPA3875	A	conserved unknown protein
RPA3956	A	ferredoxin
RPA4067	A	hypothetical protein
RPA4102	A	putative transcriptional regulator
RPA4137	A	conserved unknown protein
RPA4344	A	hypothetical protein
RPA4574	A	hypothetical protein
RPA4612	A	ferredoxin 2[4Fe-4S] III, fdxB
RPA4738	A	possible two-component system reponse regulator
RPA4836	A	30S ribosomal protein S20
RPA1697	C	Competence-damaged protein
RPA2649	C	conserved unknown protein

Table 7.6: Continued

Gene	Second AA	Function
RPA1279	D	hypothetical protein
RPA1615	D	putative methyltransferase
RPA1812	D	conserved hypothetical protein
RPA1824	D	unknown protein
RPA1842	D	conserved unknown protein
RPA2241	D	conserved hypothetical protein
RPA3130	D	Helix-turn-helix motif
RPA3524	D	putative cell division protein FtsQ
RPA3721	D	possible ABC transporter, permease protein
RPA4138	D	conserved unknown protein
RPA4191	D	conserved unknown protein
RPA4297	D	putative aldose reductase
RPA2533	E	unknown protein
RPA4179	F	conserved unknown protein
RPA0207	G	unknown protein
RPA0501	G	BolA-like protein
RPA0517	G	putative transcriptional regulator (Fur family)
RPA2648	G	unknown protein
RPA3244	G	30S ribosomal protein S3
RPA3672	G	cold shock protein
RPA3924	G	conserved hypothetical protein
RPA4418	G	conserved unknown protein
RPA0855	H	Beta-lactamase-like
RPA3123	H	hypothetical protein
RPA4605	H	electron transfer flavoprotein beta chain fixA
RPA0233	I	putative Citrate lyase beta chain (acyl lyase
RPA0600	I	conserved hypothetical protein
RPA0646	I	conserved hypothetical protein
RPA0866	I	putative nucleoside diphosphate kinase regulator
RPA1064	I	conserved hypothetical protein
RPA1108	I	Myb DNA-binding domain:DGPF domain
RPA1454	I	hypothetical protein
RPA2239	I	putative partition protein
RPA2283	I	putative proteic killer suppression protein
RPA3589	I	conserved hypothetical protein
RPA3676	I	putative type IV prepilin peptidase, cpaA
RPA3828	I	Helix-turn-helix motif
RPA3878	I	conserved unknown protein
RPA4070	I	possible peptide methionine sulfoxide reductase
RPA4542	I	unknown protein
RPA4610	I	Protein of unknown function, HesB/YadR/YfhF
RPA0885	K	conserved hypothetical protein
RPA1019	K	possible transcriptional activator HlyU
RPA1107	K	possible transcriptional regulator
RPA2057	K	hypothetical protein
RPA2314	K	cytochrome c556
RPA2522	K	hypothetical protein
RPA2667	K	conserved unknown protein
RPA3223	K	putative alginate lyase
RPA3501	K	conserved unknown protein
RPA4544	K	conserved unknown protein
RPA4676	K	putative transcriptional regulator
RPA0160	L	possible acetyltransferases.
RPA0298	L	DUF299
RPA1017	L	Nitrogen fixation-related protein
RPA1278	L	GatB/Yqey
RPA1905	L	homologue of Rhodobacter capsulatus gene

Table 7.6: Continued

Gene	Second AA	Function
RPA1985	L	probable diacylglycerol kinase
RPA2028	L	conserved hypothetical protein
RPA2603	L	conserved hypothetical protein
RPA2934	L	conserved unknown protein
RPA3101	L	conserved unknown protein
RPA3148	L	DUF174
RPA3327	L	hypothetical protein
RPA4470	L	DUF336
RPA0331	M	possible heat shock protein (HSP-70 COFACTOR),
RPA1025	M	possible Ectothiorhodospira Vacuolata
RPA2718	M	hypothetical protein
RPA2899	M	conserved hypothetical protein
RPA3719	M	putative high-affinity branched-chain amino acid
RPA4529	M	putative arsenate reductase
RPA4634	M	hypothetical protein
RPA0571	N	two-component, response regulator
RPA0594	N	putative mutator protein mutT
RPA2045	N	biotin synthetase
RPA2165	N	chaperonin GroES2, cpn10
RPA3663	N	urease gamma subunit
RPA4050	N	unknown protein
RPA4077	N	ATPase, ParA type
RPA0039	P	50S ribosomal protein L35
RPA0283	P	putative two-component response regulator
RPA0598	P	putative glutaredoxin
RPA1302	P	unknown protein
RPA1342	P	hypothetical protein
RPA1600	P	BolA-like protein
RPA1620	P	unknown protein
RPA2456	P	possible bacterioferritin co-migratory protein
RPA2690	P	possible uracil-DNA glycosylase
RPA3168	P	possible flgaellar switch protein FliN
RPA3213	P	hypothetical protein
RPA3246	P	30S ribosomal protein S19
RPA3254	P	30S ribosomal protein S7
RPA3759	P	putative 5-carboxymethyl-2-hydroxymuconate
RPA4176	P	ribosomal protein S21
RPA4228	P	hypothetical protein
RPA4600	P	conserved unknown protein
RPA4686	P	possible ABC transporter, periplasmic amino
RPA4701	P	Protocatechuate 4,5-dioxygenase, alpha chain
RPA0090	Q	hypothetical protein
RPA0263	Q	Protein of unknown function UPF0047
RPA0653	Q	2-ketocyclohexanecarboxyl-CoA hydrolase
RPA1586	Q	putative short-chain dehydrogenase/reductase
RPA1827	Q	hypothetical protein
RPA2006	Q	putative phosphatidylserine decarboxylase
RPA2453	Q	translation peptide releasing factor RF-2
RPA2717	Q	conserved hypothetical protein
RPA3215	Q	putative nitroreductase
RPA3319	Q	hypothetical protein
RPA0092	R	conserved hypothetical protein
RPA0203	R	heme exporter protein A (heme ABC transporter
RPA0301	R	putative DNA polymerase III epsilon chain
RPA0543	R	unknown protein
RPA0767	R	PAS domain:GGDEF:PAC motif
RPA0868	R	hypothetical protein

Table 7.6: Continued

Gene	Second AA	Function
RPA1289	R	hypothetical protein
RPA1915	R	FeuP two-component system, regulatory protein
RPA2523	R	putative lactoylglutathione lyase
RPA3024	R	unknown protein
RPA4277	R	conserved hypothetical protein
RPA4349	R	conserved unknown protein
RPA4689	R	conserved hypothetical protein
RPA0054	S	putative small heat shock protein
RPA0282	S	possible transcriptional regulator
RPA0433	S	ribosomal protein S15
RPA0435	S	putative ribosome-binding factor A
RPA0443	S	possible transcriptional regulator
RPA0450	S	ferric uptake regulation protein
RPA0493	S	50S ribosomal protein L28
RPA0607	S	putative protoporphyrinogen oxidase, hemK
RPA0609	S	conserved hypothetical protein
RPA0714	S	bifunctional cobinamide kinase, cobinamide
RPA0917	S	Transcriptional Regulator, AraC Family
RPA0930	S	possible 3-octaprenyl-4-hydroxybenzoate
RPA0932	S	conserved unknown protein
RPA0993	S	possible alpha-ribazole-5`-phosphate phosphatase
RPA1106	S	conserved hypothetical protein
RPA1228	S	putative 2-oxoglutarate ferredoxin
RPA1659	S	conserved unknown protein
RPA1726	S	putative oxidoreductase
RPA1757	S	possible oxoacyl carrier protein reductase
RPA1964	S	hypothetical protein
RPA2066	S	putative nosX
RPA2082	S	putative uroporphyrin III methylase
RPA2732	S	conserved hypothetical protein
RPA2744	S	hypothetical protein
RPA3073	S	constitutive acyl carrier protein
RPA3235	S	50S ribosomal protein L6
RPA3236	S	30S ribosomal protein S8
RPA3269	S	50S ribosomal protein L7/L12
RPA3290	S	possible transcriptional regulator, TetR family
RPA3394	S	DUF37
RPA3476	S	possible energy transducer TonB
RPA3555	S	arsenate reductase
RPA3852	S	hypothetical protein
RPA3896	S	hypothetical protein
RPA3913	S	conserved hypothetical protein
RPA4272	S	conserved unknown protein
RPA4383	S	conserved unknown protein
RPA4501	S	phnA-like protein
RPA4724	S	putative uridine 5-monophosphate synthase
RPA0052	T	putative nitrogen regulatory IIA protein(enzyme
RPA0059	T	L-carnitine dehydratase/bile acid-inducible
RPA0222	T	Beta-Ig-H3/Fasciclin domain
RPA0267	T	possible thioredoxin
RPA0292	T	chromosome partitioning protein, ParA
RPA0326	T	DUF24, predicted transcriptional regulator,
RPA0354	T	putative pts system phosphocarrier protein HPr
RPA0643	T	conserved hypothetical protein
RPA1168	T	molybdopterin converting factor, subunit 2
RPA1589	T	30S ribosomal protein S4
RPA1593	T	Thioesterase superfamily

Table 7.6: Continued

Gene	Second AA	Function
RPA1870	T	possible transcriptional regulator (MarR/EmrR
RPA1978	T	molybdenum biosynthetic protein A
RPA1993	T	possible virulence-associated protein
RPA2032	T	acetolactate synthase (small subunit)
RPA2084	T	precorrin 3 or 4 methylase
RPA2158	T	hypothetical protein
RPA2338	T	unknown protein
RPA2520	T	hypothetical protein
RPA2540	T	3-hydroxy-3-methylglutaryl-CoA lyase
RPA2766	T	Phenylacetic acid degradation-related
RPA2856	T	Protein of unknown function, HesB/YadR/YfhF
RPA2973	T	hypothetical protein
RPA2985	T	conserved unknownl protein
RPA3626	T	conserved unknown protein
RPA3770	T	conserved unknown protein
RPA3970	T	putative
RPA3988	T	putative phosphatase
RPA4010	T	putative response regulator
RPA4206	T	D-beta-hydroxybutyrate dehydrogenase
RPA4241	T	CBS domain
RPA4348	T	conserved hypothetical protein
RPA4357	T	conserved unknown protein
RPA4457	T	putative sulfide dehydrogenase
RPA4500	T	hypothetical protein
RPA4541	T	DNA invertase gene rlgA
RPA4602	T	ferredoxin like protein, fixX
RPA4666	T	carbon-monoxide dehydrogenase small subunit
RPA4678	T	possible outer membrane protein OprF (AF117972)
RPA4760	T	unknown protein
RPA0490	V	conserved hypothetical protein
RPA1535	V	cytochrome c2
RPA2265	V	conserved hypothetical protein
RPA2848	V	possible sec-independent protein secretion
RPA2933	V	conserved hypothetical protein
RPA3824	V	conserved hypothetical protein
RPA4072	V	transcriptional elongation factor greA
RPA4467	V	putative sulfur oxidation protein soxY
RPA4483	V	possible signal transducer
RPA4770	Y	DUF525

Table 7.7: Identification of unknown proteins with PTMs from the anaerobic growth state.

Protein	Putative PTM	Function
RPA2334	Methionine Truncation	Unknown
RPA2335	1, 2, 4 Methylations	Unknown
RPA2336	1 Methylation	Unknown
RPA2338	Methionine Truncation	Unknown
RPA1495	Methionine Truncation	Unknown
RPA1620	Methionine , 1 Methylation	Unknown

putative cation transport ATPase but does not have the predicted transmembrane domains generally associated with such a transport ATPase.

When this operon was examined with top-down methods a series of PTMs including methylations and N-terminal methionine truncations were identified [Table 7.7]. The proteins RPA2334 and RPA2338 were identified with an N-terminal methionine truncation. Interesting though, RPA2335 was identified with a series of 1-4 methylations as well as in its native form and RPA2336 was identified with 1 methylation, as seen in Figure 7.2. This unique hypothetical operon with its series of PTMs may provide a target for future functional studies such as gene knockouts and protein interaction studies through tagging protocols or other biochemical enrichment techniques. Also identified within the anaerobic growth state were the unknown proteins RPA1495 with an N-terminal methionine truncation and RPA1620 with an N-terminal methionine truncation as well as 1 methylation. Protein RPA1495 is found within an operon with light harvesting proteins which may provide a possible associated function for this protein. Methylation is a common PTM found on lysine and arginine mainly. These two residues have very polar side chains that are positively charged. When these residues are blocked by a methylation or acetylation the basic nature of that site within the protein can be changed, thereby making it more or less accessible to other targets. Within the aerobic growth state a number of proteins were identified with PTMs. Of the 426 proteins identified 394 of these possessed some form of a PTM. Included in the list of proteins that contain PTM unknown as well as common proteins such as ribosomes were identified. Two unknown proteins were of particular interest due to multiple isoforms being present.

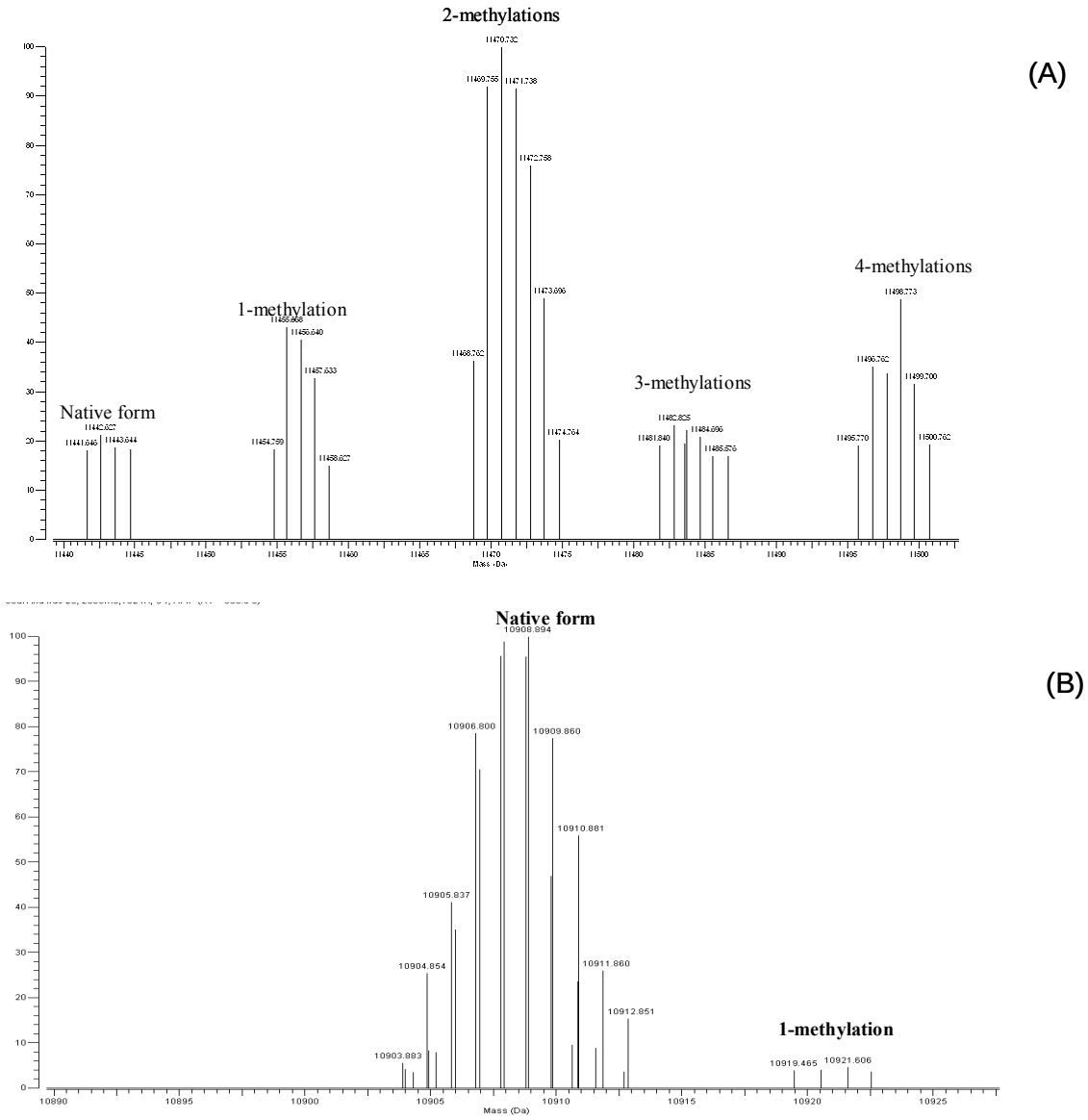


Figure 7.2: Mass spectra of RPA2335 and RPA2336. (A) Mass spectrum of RPA2335 from a unique anaerobic unknown operon showing the native, 1 methylation, 2 methylations, and 3 methylations isoforms. (B) Mass spectrum of RPA2336 from a unique anaerobic unknown operon showing the native protein and isoform with 1 methylation.

The unknown protein RPA4610 (17% sequence coverage) was found to have an N-terminal methionine truncation with a combination of 4-8 methylations [Figure 7.3]. The protein was only identified in this highly modified state making it an interesting candidate for further functional studies. Another unknown protein identified with multiple isoforms was RPA4330 which has a native form as well as a methylated version. Protein RPA0501 that was identified only in this growth state was shown to have an N-terminal methionine truncation. Two ribosomal proteins L30 and L23 also were identified with a native or an N-terminal methionine truncation and containing one methylation.

The nitrogen fixing growth state has 214 identified proteins; of these 192 have a PTM. Several of the unknown and hypothetical proteins within the nitrogen fixing growth state contain PTMs and multiple isoforms. Of particular interest are three of these conserved hypothetical proteins including, RPA2732 identified with an N-terminal methionine truncation form as well as an isoform with 2 acetylations and 1 methylation. A set of hypothetical proteins were identified within one mass spectrum from the LC-FTICR-MS data, as seen in Figure 7.4. Within this mass spectrum the first pair of proteins is RPA 1286 containing a unmodified form and a methylated isoform; the second pair are RPA2979 with an N-terminal methionine truncation plus 2 methylations and an isoform with an N-terminal methionine truncation plus 3 methylations.

A specialized PTM of interest associated with the nitrogen fixing growth state was uridylylation found on the GlnK and GlnB proteins. The GlnK and GlnB proteins are members of the pII signal transduction protein family. In *R. palustris* there are three annotated forms of pII proteins; GlnK1, GlnK2, and GlnB.

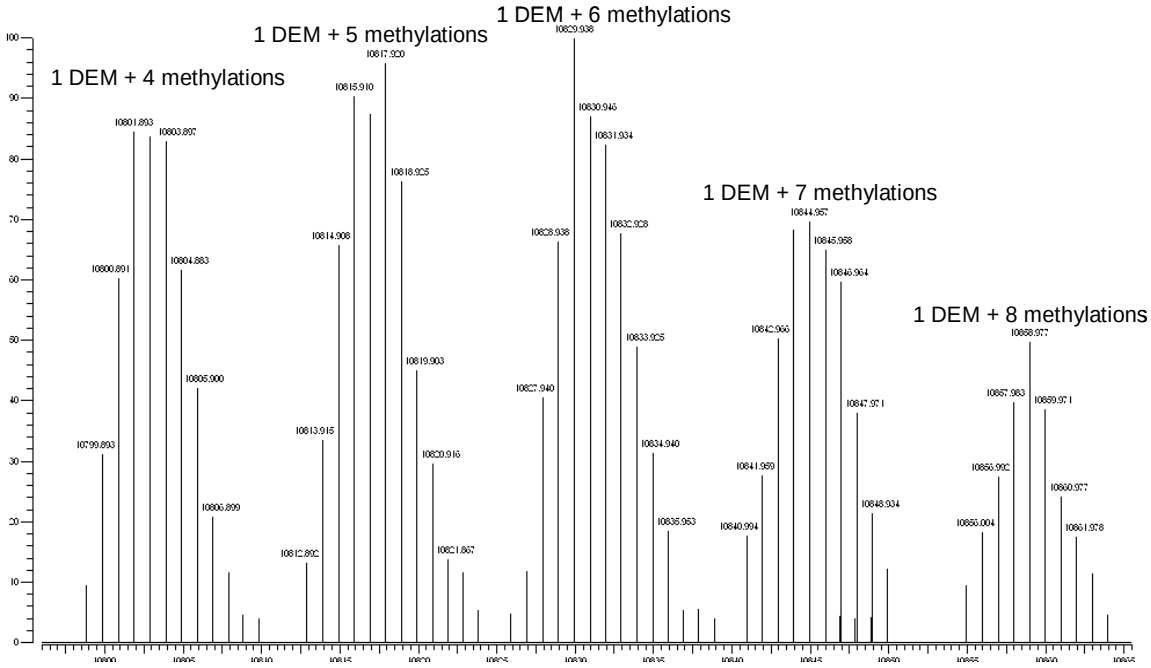


Figure 7.3: Mass spectrum of unknown protein RPA4610. The unknown protein RPA4610 with an N-terminal methionine truncation and a combination of 4-8 methylations from the aerobic growth state. DEM represent N-terminal methionine truncation within the figure.

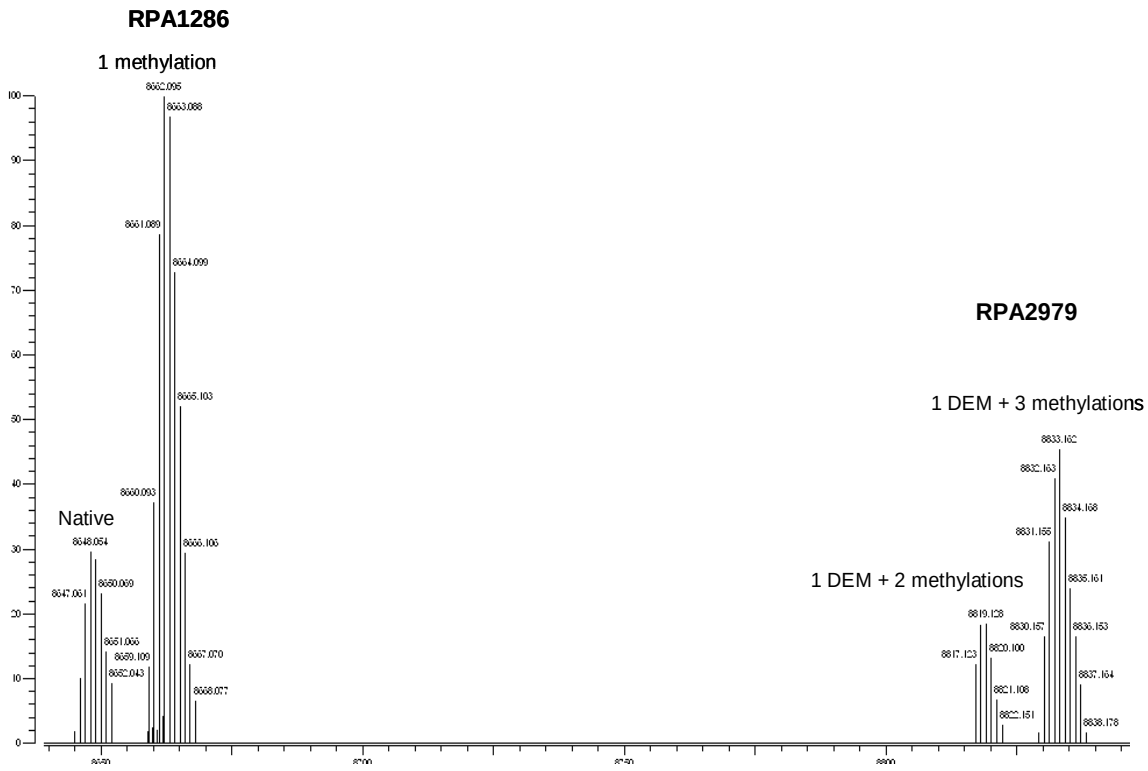


Figure 7.4: A set of hypothetical proteins identified within one mass spectrum from the LC-FTIC-MS data. Within the mass spectra the first pair of proteins is RPA1286 containing a native form and a methylated isoform; the second pair is RPA2979 with an N-terminal methionine truncation plus 2 methylations and an isoform with an N-terminal methionine truncation plus 3 methylations.

Two of these proteins were found to undergo uridylylation under ammonium starvation conditions (nitrogen fixing), presumably to regulate the AmtB ammonium transporter as well as glutamine synthetase [102]. Under nitrogen fixing growth conditions the GlnK2 (RPA0274) and GlnB (RPA2966) proteins were identified in both the unmodified and modified states [123]. The GlnK2 protein was identified in the nitrogen fixing growth state along with GlnB, while GlnK1 was identified in the anaerobic growth states. These are the states the proteins should be found in according to previous research (Chapter 5).

Signal Peptides

Top-down mass spectrometry can provide information on the function and location of proteins. This is especially true when proteins containing signal peptides are considered. Most cell types and organisms employ several ways of targeting proteins to the extracellular environment or subcellular locations. Most of the proteins targeted for the extracellular space or subcellular locations carry specific sequence motifs (signal peptides) characterizing the type of secretion/targeting it undergoes. To identify potential amino-terminal signal peptides, primary sequence analysis of the predicted *R. palustris* proteome was performed by the SignalP NN [124], SignalP HMM [124], PrediSi [125], and PSORTb [126] algorithms. A subdatabase containing *R. palustris* proteins with predicted signal peptides by all three signal peptide prediction algorithms was created by Judson Hervey, a graduate student in the Genome Science and Technology program. Amino-terminal signal peptides were removed from each protein in the subdatabase based upon the predicted cleavage site by the SignalP NN(2) algorithm. Within the three growth states of *R. palustris* examined, 22 proteins with predicted signal peptides were identified using the database of predicted proteins containing signal peptides [Table 7.8].

Table 7.8: Identified proteins with signal peptides.

Protein	Unprocessed MW	Processed MW	Measured Mass	Sequence Coverage	Function
RPA0088	7665.972	5687.447	5687.404		Unknown Protein
RPA0090	15280.276	8049.5894	8048.786	7.5	Hypothetical Protein
RPA0091	11250.928	7572.476	7572.525		Hypothetical Protein
RPA0744	9858.252	6185.924	6186.192		Putative High Potential Iron Sulfur Protein
RPA1023	11434.053	8706.657	8706.918		Hypothetical Protein
RPA1088	12609.711	8223.414	8223.715	13.5	Hypothetical Protein
RPA1428	27812.97	25962.713	25962.602	19.8	Possible Lipoprotein
RPA1454	9442.541	6067.368	6067.362	21.6	Hypothetical Protein
RPA1824	17579.16	12571.359	12571.258	8.2	Unknown Protein
RPA1847	9430.428	7119.5864	7120.332		Conserved Hypothetical Protein
RPA1874	7297.609	5526.5815	5526.812		Hypothetical Protein
RPA2544	8676.271	5865.847	5865.694		Conserved Hypothetical Protein
RPA2546	15826.283	12707.527	12707.453	34.9	FKBP-type Peptidyl-prolyl cis-trans Isomerase
RPA3025	11292.242	7926.262	7926.935		Hypothetical Protein
RPA3034	9315.853	6270.1743	6270.788	19.8	Unknown Protein
RPA3101	17265.908	13337.184	13337.205	8.8	Unknown Protein
RPA3362	10279.998	7755.001	7754.766		Unknown Protein
RPA3373	10172.86	8148.3096	8149.009	15.5	Hypothetical Protein
RPA3957	12532.322	6725.5767	6725.496		Hpt Domain
RPA4329	15828.192	13645.581	13645.413	9	Unknown Protein
RPA4467	16279.965	12802.759	12802.356	9.8	SoxY2 Putative Sulfur Oxidation Protein
RPA4573	16557.592	15352.109	15351.637	27.8	Unknown Protein

Twelve of the 22 identified proteins have some bottom-up sequence coverage. However, the signal peptide is not able to be identified by bottom-up methods the rest of the protein can provide peptide information for identification. The remaining 10 proteins identified that do not have bottom-up sequence coverage are generally too small to be detected, in bottom-up, after truncation In the case of signal peptide searching the union of top-down and bottom-up identifications are shown, as well as proteins identified with only top-down searching [Table 7.8].

Seventeen of the proteins identified were unknown or hypothetical proteins. The identification of signal peptides from these proteins provides a basis for starting to determine the function and location of these proteins. The putative high potential iron-sulfur protein (RPA0744), FKBP-type peptidyl-prolyl cis-trans isomerase (RPA2546), and soxY2 putative sulfur oxidation protein (RPA4467) were all proteins that were identified with known functions. These three proteins have functions that a signal peptide would expect to been seen for. For example, the putative high potential iron-sulfur proteins are a specific class of high-redox potential 4Fe-4S ferrodoxins that function in anaerobic electron transport and which occurs in photosynthetic bacteria. Also, this protein has been shown to have predicted signal peptides in other bacteria. In *R. palustris* the putative high potential iron-sulfur protein (RPA0744) was identified in the anaerobic nitrogen fixing growth state, which correlates with its function in electron transport during anaerobic growth. The integrated top-down and bottom-up approach provided for the identification of 22 signal peptides in R. palustris, which gives an additional level of information about this organism.

Conclusions

In this study, we have characterized the *R. palustris* proteome by integrated top-down and bottom-up analysis under three major metabolic states. We confidently identified 599 proteins by an integrated top-down and bottom-up approach. In total, 241 proteins classified as unknown and conserved unknown proteins were identified, representing 17.3% of the identified proteins. Over 500 proteins were identified containing some form of a PTM. The proteome analysis of a number of metabolic states with their associated PTMs and isoforms is necessary to begin to understand how microbes change their proteome to adapt to the resources present. The conserved unknown and unknown proteins that were identified as containing multiple isoforms under the metabolic states examined here are excellent targets for future studies, because they may have important functions under those states. The detection of PTMs on an unknown operon of five proteins found to be expressed only under the phototrophic (anaerobic illuminated) states, with no evidence of expression under chemotrophic (aerobic dark) states, was an excellent example of the discovery capabilities of this general method to provide further information of function and location for these proteins.

Our data indicates that it is possible to identify large numbers of intact proteins with and with out PTMs and correlate this information to bottom-up ms/ms data. By creating a list of common PTMs one can begin the process of imparting information about the natural state of the protein and how it may be functioning within the cell. This is the first study of this magnitude to offer such a comprehensive list of intact identified proteins with their associated PTMs. This employed technique should provide a starting point of future work with protein complexes and functional studies within this as well as other microbial systems.

Chapter 8

Conclusions and Impact of Integrated and Computational Platform for the Analysis of Intact Proteins and PTMs of Microbial Systems by Top-down Mass Spectrometry

The overall goal of this dissertation research was to develop an integrated computational and experimental platform for characterizing protein isoforms and PTMs in microbial systems by top-down FT-ICR mass spectrometry. We first evaluated the methodologies of microbial growth, intact protein and protein complex extractions, followed by sample preparation and then progressed to identification of the instrumentation needed to integrate the two methodologies used in these studies. Emphasis was placed on the development of integrated top-down and bottom-up informatics and the challenges faced in the integration of these two large data sets and extraction of relevant biological data. We then illustrated how these technologies can be applied to the analysis of complex protein mixtures, protein complexes and microbial proteomes. Great progress has been made through these studies, but much work is still needed in the areas of intact protein separations, data data-dependent MS/MS on intact proteins within liquid chromatography time scales, and further analysis of PTMs once tentatively identified. Some avenues of research performed in this dissertation to combat these issues are discussed below.

During this dissertation work, an essential need for fundamental advancements in the analysis of proteins and peptides was addressed. Two areas of particular interest included better methods of determination of charge states for large proteins and advanced protein fragmentation methods with the FTICR-MS. Each of these areas was addresses in this dissertation work. Due to the difficulties encountered with LC-FTICR-MS

measurements and charge state determination, an automated method for determining charge states from high-resolution mass spectra was developed. Fourier transforms of isotope packets from high-resolution mass spectra are compared to Fourier transforms of modeled isotopic peak packets for a range of charge states. The charge state for the experimental ion packet is determined by the model isotope packet that yields the best match in the comparison of the Fourier transforms. Existing charge state assignment algorithms for FTICR-MS data appear to require centroiding before charge determination, and errors in this process can lead to errors in assessed charges. Use of Fast Fourier transforms (FFT) for charge determination does not require centroiding and appears to achieve superior sensitivity and noise suppression than algorithms of this type, especially for LC-FTICR-MS measurements. This advancement can be applied to data analysis in order to ensure the most accurate protein identifications during searching against a protein database. The second area targeted for FTICR-MS development was the evaluation of proteins and peptide fragmentation methods within the FTICR-MS. This work demonstrated the use of MSAD as a replacement for more commonly applied fragmentation methods, such as SORI-CAD, within the FTICR as a feasible option for simple peptides solutions, tryptic digest and simple mixtures. MSAD provides a fragmentation method that can fragment all peptides in the sample in one step eliminating the isolation step needed for SORI-CAD, which provides a more operationally simple and time saving method. MSAD saves time during the experimentation process, although, the data analysis is in-depth and time consuming due to the complexity of the fragmentation spectra. Therefore, at this time we are not employing MSAD for intact protein analysis. These two fundamental studies provided better methods for protein

charge state determination under liquid chromatography conditions. Also, these studies provided alternative fragmentation methods of proteins and peptides within the FTICR-MS, thereby advancing the field of top-down mass spectrometry.

The combination of the top-down and bottom-up MS methodologies for the characterization of individual proteins, protein complexes and whole proteomes were the major focus of this dissertation work. While many proteomics groups are focusing on either top-down or bottom-up techniques, very few have tried to integrate the two technologies. Through the work of this dissertation we have pushed the forefront of this technology. Our initial effort was to analyze complex ribosomal protein mixtures from *R. palustris* and antibiotic resistant *E. coli* strains; this effort showed great promise for this integrated technology to obtain a detailed level of information not possible by either technique alone. This includes the determination of the position and number of post-translational modifications on the intact protein product, as well as the determination of the number and position of amino acid changes (mutations) within intact proteins for most potential substations (Ile-Leu can't be resolved because they are isobaric). The integrated top-down and bottom-up analysis of component proteins of the 70S ribosome from *R. palustris* enhanced several aspects of the analysis. For this study, the bottom-up approach was expanded to the use of 1D and 2D LC-MS/MS methodologies for the analysis of the enzymatically digested ribosomal protein complex. For the experiments on *R. palustris* ribosomal complexes, we performed LC-ES-FT-ICR for intact protein measurements. Not only was this method useful in the analysis of *R. palustris* ribosomes; the use of integrated top-down and bottom-up mass spectrometry approaches provided insight into the role of ribosomal proteins in streptomycin resistance in *E. coli.* In this

study, we employ an integrated top-down and bottom-up approach to characterize the ribosomal proteins from wild type K12 and two streptomycin resistant strains of *E. coli*. Using this method, a complement of ribosomal proteins with unique PTM series, isoforms, and point mutations were identified from all three strains. For the first time, this method allowed for the interrogation of differential post translational modifications in the "compensation" process for *E. coli*, as well as further conformation of point mutations thought to confer antibiotic resistance.

The analysis of key regulation sites within protein complexes was the next step in the development of the integrated top-down and bottom-up platform. To perform this analysis, affinity purifications of the *R. palustri*s pII family of proteins consisting of GlnK1, GlnK2 and GlnB were analyzed. In bacteria, the pII family generally plays a pivotal role in nitrogen metabolism regulation due to its ability to sense internal cellular ammonium concentrations. The uridylylation of these proteins regulate ammonia transporters as well as glutamine synthetase. Affinity purifications in conjunction with top-down and bottom-up mass spectrometry permitted the isolation and characterization of the functional state and isoforms for these proteins as a function of nitrogen availability. From this work, it was determined that under nitrogen fixing conditions, all of these pII proteins are uridylylated, all on the Tyr-51 positions. Thus, pII protein uridylylation appears to be tightly coordinated with nitrogen availability. By using a combined technique of protein affinity purifications and mass spectrometry, it was determined, for the first time, that GlnK2, GlnK1 and GlnB proteins possess an uridylylation under nitrogen fixing growth conditions in *R. palustris*. This information allowed for a previously un-afforded glimpse into the modifications and isoforms of the

proteins that regulate the AmtB transporter and glutamine synthetase in *R. palustris*. Not only did this method provide a glimpse into a key regulation site for a protein complex it also expanded the capabilities of this approach for future systems.

At the outset of this dissertation a primary limitation of top-down analysis was bioinformatics tools for querying protein databases. The isotopic packets of intact proteins and the MS/MS spectra of intact proteins are both much more complicated than those derived from peptide measurements thereby enhancing this problem. This dissertation work provided the first informatics tools for combining top-down and bottom-up datasets to search for PTMs, amino acid substitutions, and N-terminal truncations. At the start of this dissertation, the ProSight PTM and PROCLAIM algorithms had been available for the analysis of intact protein and their MS/MS spectra against protein databases as well as PTM prediction. Even with these programs, no major effort had been made to integrate top-down analysis with traditional enzymatic bottom-up analysis for protein identification and PTM analysis. Our ORNL developed algorithm PTMSearchPlus is the first software providing a comprehensive search method that allows for the integration of top-down protein identification with the bottom-up peptide data to identify proteins and their associated PTMs. The software is built around multiple instrumentation platforms and data inputs. These multiple instrumentation and data platforms include bottom-up ion trap data, as well as top-down high resolution data such as FTICR data. The software can accomplish independent top-down or bottom-up searches, as well as these two parts of the program being able to interact in a combined search. By combining these two search capabilities, the results from the top-down search can limit the number of the proteins that are used to generate the database used for the

bottom-up search (search time decrease) and in return, the results of the bottom-up search can be used as a confirmation for the proteins with associated PTMs found in the top-down search. This integration reduces the search time dramatically, allowing the user to search for more PTMs on proteins and peptides during a reasonable time frame. The software was demonstrated with a protein standard mixture and complex ribosomal protein mixture. All proteins from the protein standard mixture, which was used as a training set, were identified using PTMSearchPlus. The *R. palustris* complex ribosomal mixture was previously examined in an integrated fashion by manual comparison. Using PTMSearchPlus all of the identified ribosomal proteins identified in the previous study were identified in a fraction of the time. Both of these test cases showed the power of the integrated approach as well as demonstrating the accuracy and speed of PTMSearchPlus.

The final goal of this dissertation was to apply the developed integrated computational and experimental platforms developed to intact proteomes of microbial systems under different growth conditions. This is the first study of this magnitude to offer such a comprehensive list of intact identified proteins with their associated PTMs. Within this study, the first large-scale characterization of three growth states of *R. palustris* by an integrated top-down and bottom-up approach was performed. This global measurement strategy was able to provide information on intact proteins, including PTMs, isoforms, and signal peptides from a given growth state. The technological approaches developed in this dissertation provided information on the function and location of proteins, as well as providing confirming peptide MS/MS data. These tools were shown to be especially powerful when determining what modification states play a role in the switch between different growth conditions, characterizing known and

unknown proteins, and determining trends within protein expression across the chosen metabolic states. Our data indicates that it is possible to identify large numbers of intact proteins with and without PTMs and correlate this information to bottom-up MS/MS data. This technique should provide a starting point of future work with protein complexes and functional studies within this, as well as other microbial systems.

This dissertation provided the first comprehensive platform for integrated top-down and bottom-up analysis of proteins, but many areas of work remain. Four of the most important areas of work include intact protein separations, data-dependent MS/MS on intact proteins within liquid chromatography time scales, intact protein bioinformatics, and further analysis of PTMs once tentatively identified. Top-down technology in its current form has difficulties with the complex mixtures found in whole proteome analysis. Potential 2D separations of intact proteins such as the off-line FPLC followed by on-line HPLC employed in this dissertation may overcome some of these limitations by providing less complex protein fractions to analyze. However, the loss of protein is always a concern when employing multiple protein purification and separation steps, this is necessary to reduce the protein complexity from a proteome into more manageable fractions for the mass spectrometer. The main area of concern, though, is the inability to separate some protein sizes and types with the commonly employed C4 reverse phase chromatography, such as proteins larger than 40- 50 kDa. This limitation exists due to on-line chromatography of intact proteins is often difficult; because of the wide range of protein sizes and hydrophobicities within the complex proteome mixtures. Parts of this problem can be addressed by employing shorter carbon chain reverse phase columns, such as a C2 column. Another option is to use different stationary phases for intact

protein separation. One example of an alternative is hydrophilic interaction chromatography (HILIC). HILIC is a variant of normal phase chromatography, where the stationary phase must be extremely polar. The elution order with HILIC is least to most polar, the opposite of that in reverse phase liquid chromatography. This method provides promise, although, there are still issues with protein precipitation that need to be worked out. One other option is buffer additives, such as hexafluroisoporponal (HFIP), which acts as a chaotrope to help provide better separations. These three solutions do provide some benefit, but in the future better separation methods for intact proteins are needed.

One of the primary technological advances needed for this combined technology includes methods for data-dependent MS/MS on intact proteins on liquid chromatography time scales. Currently, the methods of IRMPD and ECD are employed for intact protein MS/MS. These methods, while powerful, still leave room for improvement in the ability to perform them on a liquid chromatography time scale and accomplish extensive fragmentation of the protein. New instrumentation, such as the use of resolving quadrupoles within the FT-ICR may help. Using the resolving quadrupole in the front of the FT-ICR the proteins can be targeted for dissociation more readily. Also the use of IRMPD in tandem with ECD provides two distinct forms of fragmentation and provides a wider range of fragmentation for proteins of varying sizes. Hopefully the continuation of fundamental instrumentation research will provide some of the answers to this limitation.

Work in this dissertation moved the field of top-down bioinformatics forward, by using an integrated top-down and bottom-up search method found in PTMSearch Plus. This program provides a great advance in integrated top-down and bottom-up searching, but more work is need in the areas of addressing point mutation, signal peptides, and

truncations. Amino Acid point mutations in proteins are one of the most difficult areas to include in a protein identification algorithm, due to the enormous combination of possibilities for all of the 20 amino acids within the protein sequence. Another area of future work is the automated prediction and identification of signal peptides. Signal sequences play an important role in protein processing and identification can sometimes be crucial to determining a protein's function and possible location within the cell (i. e. if located within the periplasm). Hopefully, future work will advance the area of intact protein analysis by providing quick and automated ways of identifying these PTMs.

The integrated top-down and bottom-up technology has already allowed for the characterization of hundreds of conserved unknown and unknown proteins and their associated PTMs as well as PTMs on known proteins. One of the clearest challenges is the integration of the field of PTM analysis in proteomics with rapid structural analysis, functional assays and genetic methods to develop rapid integrated methods to determine not only the identity of conserved unknown and unknown proteins, but also their function and the role the PTMs on them play. Another challenge is to determine what role identified PTMs play in the regulation of known proteins. For top-down proteomics to become truly useful at gaining insight into the function and regulation of the many proteins present in any microbial species, this must be accomplished

The final major challenge is the application of this technology to microbial growth states for rapid and routine analysis. While small important steps were taken in the course of this dissertation, much work is still needed. The complexity of PTMs on proteins is truly daunting but unless initial steps are taken to attack this important aspect no progress will be made. Hopefully, the work presented in this dissertation brings us one

step closer to the ultimate goal of an integrated computational and experimental platform for characterizing protein isoforms and PTMs in microbial systems by top-down FT-ICR mass spectrometry.

List of References

[1] Ideker, T.; Galitski, T.; Hood L. (2001) A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet*., **2**, 343-372.

[2] James, P. (2001) Mass Spectrometry and the Proteome, in *Proteome Research: Mass Spectrometry*, P. James, editor, Springer, Germany, p.6.

[3] Alberts, B.; Bray, D.; Lewis, J.; Raff, M.; Roberts, K. and J. D. Watson, editors. (1994) "Molecular Biology Of the Cell," Garland Publishers, New York. (ISBN 0-8153-1619-4)

[4] Driessen, H. P. C.; de Jong, W. W.; Tesser, G. I.; Bloemendal, H. (1985) The mechanism of amino-terminal acetylation of proteins. *CRC Crit Rev Biochem*. **18**, 281–325.

[5] Persson, B.; Flinta, C.; von Heijne, G.; Jörnvall, H. (1985) Structures of amino-terminally acetylated proteins. *Eur J Biochem*. **152**, 523–527.

[6] Polevoda, B.; Sherman F. (2000) N^{α}-terminal acetylation of eukaryotic proteins. *J Biol Chem*. **275**, 36479–39482.

[7] Polevoda, B; Norbeck, J.; Takakura, H.; Blomberg, A.; Sherman F. (1999) Identification and specificity of amino-terminal acetyltransferases from *Saccharomyces cerevisiae*. *EMBO J*. **18**, 6155-6168.

[8] Hunter, T. (1987) A thousand and one protein kinases. *Cell.* **50**, 823-829.

[9] Kim, J. and Kendall, D. A. (2000) Cell stress chaperones abstracts. *Cell Stress Chaperones*. **5(4)**, 267-275.

[10] Harrison, F. H. and Harwood, C. S. (2005) The p*imFABCDE* operon from *Rhodopseudomonas palustris* mediates dicarboxylic acid degradation and participates in anaerobic benzoate degradation. *Microbiology*. **74**, 727-736.

[11] Harwood, C. S. and Gibson. (1988) Anaerobic and aerobic metabolism of diverse aromatic compounds by the photosynthetic bacterium *Rhodopseudomonas palustris. J. Appl. Environ. Microbiol*. **54**, 712-717.

[12] Samanta, S. K.; Harwood, C. S. (2005) Use of the *Rhodopseudomonas palustris* genome to identify a single amino acid that contributes to the activity of a coenzyme A ligase with chlorinated substrates. *Mol. Microbiol*. **55**, 1151- 1159.

[13]Larimer, F. W. et al. (2004) Complete genome sequence of the metabolically versatile photosynthetic bacterium *Rhodopseudomonas palustris. Nature Biotech*. **22**, 55-60.

[14] Oda, Y.; Samanta, S. K.; Rey, F. E.; Wu, L.; Liu, X.; Yan, T.; Zhou, J.; Harwood: C. S. J. (2005) Use of Rhodopseudomonas palustris genome sequence to identify a single amino acid that contributes to the activity of a coenzyme A ligase with chlorinated substrates.*Bacteriology,* **187**, 7784-7794.

[15]Buchanan, M.V. et al. (2002) Genomes to Life "Center for Molecular and Cellular Systems": a research program for identification and characterization of protein complexes. *OMICS*, **6**, 287-303.

[16] Blattner, F. R. et al. (1997). The complete genome sequence of Escherichia coli K-12. *Science*, **277**,1453-1462.

[17] Fenn, J.B.; Mann, M.; Meng, C.K.; Wong, S.F.; Whitehouse, C.M. (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**, 64-71.

[18] Hillenkamp, F.; Karas, M.; Beavis, R.C.; Chait, B.T. (1991) Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Anal. Chem.* **63**, 1193A-1203A.

[19] Nakanishi T.; Okamoto N.,; Tanaka, K.,; Shimizu, A. (1994) Laser-Desorption Time-Of-Flight Mass-Spectrometric Analysis Of Transferrin Precipitated With Antiserum - A Unique Simple Method To Identify Molecular-Weight Variants. *Biological Mass Spectrometry* **23,** 230-233.

[20] Peng J. and Gygi, S.P. (2001) Proteomics: The Move to Mixtures. *J. Mass Spec*. **36**, 1083-1091.

[21] Larsen, M.R. and Roepstorff, P. (2000) Mass spectrometric identification of proteins and characterization of their post-translational modifications in proteome analysis. *Fresenius J. Anal. Chem*. **366**, 677-690.

[22] Little, D.P.; Speir, J.P.; O'Connor, P.B.; McLafferty, F.W. (1994) Infrared multiphoton dissociation of large multiply charged ions for biomolecule sequencing. *Anal. Chem.* **66**, 2809-2815.

[23] Mortz, E.; O'Connor, P.B.; Roepstorff, P.; Kelleher, N.L.; Wood, T.D.; McLafferty, F.W. Mann, M. (1996) Sequence tag identification of intact proteins by matching tandem mass spectral data against sequence data bases. *Proc. Natl. Acad. Sci. USA* **93**, 8264-8267.

[24] Kelleher, N.L.; Taylor, S.V.; Grannis, D.; Kinsland, C.; Chiu, H.J.; Begley, T.P.; McLafferty, F.W. (1998) Efficient sequence analysis of the six gene products (7-74 kDa) from the *Escherichia coli* thiamin biosynthetic operon by tandem high-resolution mass spectrometry. *Protein Sci*. **7**, 1796-1801.

[25] McLuckey, S.A. and Stephenson, J.L. Jr. (1998) Ion/ion chemistry of high-mass multiply charged ions. *Mass Spectrom. Rev.* **17**, 369-407.

[26] Hess, D.; Covey, T.C.; Winz, R.; Brownsey, R.W.; Aebersold, R. (1993) Analytical and micropreparative peptide mapping by high performance liquid chromatography/electrospray mass spectrometry of proteins purified by gel electrophoresis. *Protein Sci.* **2**, 1342-1351.

[27] Mortz, E.; Vorm, O.; Mann, M.; Roepstorff, P. (1994) Identification of proteins in polyacrylamide gels by mass spectrometric peptide mapping combined with database search. *Biol. Mass Spectrom.* **23**, 249-261.

[28] Shevchenko, A.; Jensen, O.N.; Podtelejnikov, A.V.; Sagliocco, F.; Wilm, M.; Vorm, O.; Mortensen, P.; Shevchenko, A.; Boucherie, H.; Mann M. (1996a) Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc. Natl. Acad. Sci. USA* **93**, 14440-14445.

[29] Wilm, M.; Shevchenko, A.; Houthaeve, T.; Breit, S.; Schweigerer, L.; Fotsis, T.; Mann M. (1996) Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature* **379**, 466-469.

[30] Gatlin, C.L.; Kleemann, G.R.; Hays, L.G.; Link, A.J.; Yates, J.R. 3rd (1998) Protein identification at the low femtomole level from silver-stained gels using a new fritless electrospray interface for liquid chromatography-microspray and nanospray mass spectrometry. *Anal. Biochem.* **263**, 93-101.

[31] McCormack, A.L.; Schieltz, D.M.; Goode, B.; Yang, S.; Barnes, G.; Drubin, D.; Yates, J.R. 3rd (1997) Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level. *Anal. Chem.* **69**, 767-776.

[32] Martin, S.E.; Shabanowitz, J.; Hunt, D.F.; Marto, J.A. (2000) Subfemtomole MS and MS/MS peptide sequence analysis using nano-HPLC micro-ESI fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* **72**, 4266-4274.

[33] Shen, Y.; Zhao, R.; Belov, M.E.; Conrads, T.P.; Anderson, G.A.; Tang, K.; Paša-Tolić, L.; Veenstra, T.D.; Lipton, M.S.; Udseth, H.R.; Smith, R.D. (2001) Packed capillary reversed-phase liquid chromatography with high-performance electrospray

ionization Fourier transform ion cyclotron resonance mass spectrometry for proteomics. *Anal. Chem.* **73**, 1766-1775.

[34] Kelleher, N. (2004) Top-Down Proteomics *Anal. Chem*. **76**, 196A-203A.

[35] VerBerkmoes, N.C.; Bundy, J.L.; Hauser, L.; Asano, K.G.; Razumovskaya, J.; Larimer, F.; Hettich, R.L.; Stephenson, J.L. Jr. (2002) Integrating "Top-Down" and "Bottom-Up" mass spectrometric approaches for proteomic analysis of *Shewanella oneidensis*. *J. Proteome Res.* **1**, 239-252.

[36] Blank, P.S.; Sjomeling, C.M.; Backlund, P.S.; Yergey, A.L. (2002) Use of cumulative distribution functions of characterize mass spectra of intact proteins. *J. Am. Soc. Mass Spectrom.* **13**, 40-46.

[37]Gomez, S.M.; Nishio, J.N.; Faull, K.F.; Whitelegge, J.P. (2002) The chloroplast grana proteome defined by intact mass measurements from liquid chromatography mass spectrometry. *Mol. Cell Proteomics* **1**, 46-59.

[38] Lee, S.-W.; Berger, S.J.; Martinović, S.; Paša-Tolić, L.; Anderson, G.A.; Shen, Y.; Zhao, R.; Smith, R.D. (2002) Direct mass spectrometric analysis of intact proteins of the yeast large ribosomal subunit using capillary LC/FTICR. *PNAS.* **99**, 5942-5947.

[39] Meng, F.; Cargile, B.J.; Patrie, S.M.; Johnson, J.R.; McLoughlin, S.M.; Kelleher, N.L. (2002) Processing complex mixtures of intact proteins for direct analysis by mass spectrometry. *Anal. Chem.* **74**, 2923-2929.

[40] Merchant, M. and Weinberger, S.R. (2000) Recent advancements in surface-enhanced laser desorption/ionization-time of flight-mass spectrometry. *Electrophoresis* **21**, 1164-1167.

[41] Ogorzalek Loo, R.R.; Cavalcoli, J.D.; VanBogelen, R.A.; Mitchell, C.; Loo, J.A.; Moldover, B.; Andrews, P.C. (2001) Virtual 2-D gel electrophoresis: Visualization and analysis of the *E. coli* proteome by mass spectrometry. *Anal. Chem.* **73**, 4063-4070.

[42] Reid, G.E. and McLuckey, S.A. (2002) 'Top down' protein characterization via tandem mass spectrometry. *J. Mass Spectrom.* **37**, 663-675.

[43] Meng, F.; Cargile, B.J.; Miller, L.M.; Forbes, A.J.; Johnson, J.R.; Kelleher, N.L. (2001) Informatics and multiplexing of intact protein identification in bacteria and the archaea. *Nature Biotech.* **19**, 952-957.

[44] Stephenson, J.L. Jr.; Cargile, B.J.; McLuckey, S.A. (1999) Ion Trap Collisional Activation of Disulfide Linkage Intact and Reduced Multiply Protonated Polypeptides. *Rapid Comm. Mass Spec*. **13**, 2040-2048.

[45] Nemeth-Cawley, J.F. and Rouse, J.C. (2002) Identification and sequencing analysis of intact proteins via collision-induced dissociation and quadrupole time-of-flight mass spectrometry. *J. Mass Spectrom.* **37**, 270-282.

[46] McLafferty, F.W.; Horn, D.M.; Breuker, K.; Ge, Y.; Lewis, M.A.; Cerda, B.; Zubarev, R.A.; Carpenter, B.K. (2001) Electron capture dissociation of gaseous multiply charged ions by Fourier-transform ion cyclotron resonance. *J. Am. Soc. Mass Spectrom.* **12**, 245-249.

[47] Horn, D.M.; Ge, Y.; McLafferty, F.W. (2000a) Activated ion electron capture dissociation for mass spectral sequencing of larger (42 kDa) proteins. *Anal. Chem.* **72**, 4778-4784.

[48] Ge, Y.; Lawhorn, B.G.; ElNaggar, M.; Strauss, E.; Park, J.-H.; Begley, T.P.; McLafferty, F.W. (2002) Top down characterization of larger proteins (45 kDa) by electron capture dissociation mass spectrometry. *J. Am. Chem. Soc.* **124**, 672-678.

[49] Demirev, P.A.; Ramirez, J.; Fenselau, C. (2001) Tandem mass spectrometry of intact proteins for characterization of biomarkers from Bacillus cereus T spores *Anal. Chem*. **73**, 5725-5731

[50] Laskin J.; Futrell J.H. (2003) Collisional Activation of Peptide Ions in FT-ICR Mass Spectrometry. *Mass Spectrometry Reviews*. **22**, 158-181.

[51] Forbes, A.J.; Mazur, M.T.; Patel, H.M.; Walsh, C.T.; Kelleher, N.L. (2001) Toward efficient analysis of >70 kDa proteins with 100% sequence coverage. *Proteomics.* **1**, 927-933.

[52] Li, W.; Hendrickson, C.L.; Emmett, M.R.; Marshall, A.G. (1999) Identification of intact proteins in mixtures by alternated capillary liquid chromatography electrospray ionization and LC ESI infrared multiphoton dissociation Fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* **71**, 4397-4402.

[53] Little D.P.; Speir J.P.; Senko M.W.; O'Connor P.B.; McLafferty F.W. (1994) Infrared Multiphoton Dissociation of Large Mulitiply Charged Ions for Biomolecule Sequencing. *Anal. Chem*. **66**, 2809-2815.

[54] Strader, M.B. et al. (2004) Characterization of the 70S Ribosome from *Rhodopseudomonas palustris* using an integrated "top-down" and "bottom-up" mass spectrometric approach. *J. Proteome Res*., **3**, 965-978.

[55] http://www.ornl.gov/sci/GenomestoLife/index.shtml

[56] Mann, M. and Pandey, A. (2001). Use of mass spectrometry-derived data to annotate nucleotide and protein sequence databases. *Trends Biochem Sci***. 26**, 54-61.

[57] Marshall, A. G., C. L. Hendrickson, et al. (1998). Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrom Rev***. 17**, 1-35.

[58] Henrickson, C. L. and Emmett C. R. (1999). Electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Annu Rev Phys Chem***. 50**, 517-36.

[59] Sharp, J.S; Becker, J.M.; Hettich, R.L. (2003) Protein surface mapping by chemical oxidation: structural analysis by mass spectrometry. *Anal. Biochem*., **312,** 216-225.

[60] Stafford, G. (2002) Ion Trap Mass Spectrometry: A personnel perspective. *J. Am. Soc. Mass. Spectrom.*, **13**, 589-596.

[61] Tabb, D.L.; Hayes-McDonald, W.; Yates, J.R. (2002) DTASelect and Contrast: Tools for Assembling and Comparing Protein Identifications from Shotgun Proteomics. *J. Proteome Res.*, **1**, 21-26.

[62] Eng, J.K.; McCormack, A.L.; Yates, J.R. 3rd (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Mass Spectrom.*, **5**, 976-989.

[63]Narasimhan, C.; Tabb, D. L.; VerBerkmoes, N. C.; Thompson, M. R.; Hettich, R. L.; Uberbacher, E. C.(2005) MASPIC: Intensity-based tandem mass spectrometry scoring scheme that improves peptide identification at high confidence. *Anal. Chem.* **77***,* 7581–7593.

[64] Marshall, A. G.; Hendrickson, C. L. (2002) Fourier transform ion cyclotron resonance detection: Principles and experimental configurations. *Int. J. Mass Spectrom.* **215**, 59–75.

[65] Laskin J.; Futrell J.H. (2003) Collisional Activation of Peptide Ions in FT-ICR Mass Spectrometry. *Mass Spectrometry Reviews*. **22**, 158-181.

[66] Zubarev R.A.; Kelleher N.L.; McLafferty F.W. (1998) Electron Capture Dissociation of Multiply Charged Protein Cations. *J. Am. Chem. Soc*. **120**, 3265-3266.

[67] Laskin, J.; Denisov E.V.; Shukla A.K.; Barlow S.E.; Futrell J.H.(2002) Surface-induced dissociation in a Fourier transform ion cyclotron resonance mass spectrometer: New instrument design and evaluation. *Anal Chem*. **74**, 3255-3261.

[68] Little D.P.; Speir J.P.; Senko M.W.; O'Connor P.B.; McLafferty F.W. (1994) Infrared Multiphoton Dissociation of Large Mulitiply Charged Ions for Biomolecule Sequencing. *Anal. Chem*. **66** (1994) 2809-2815.

[69] K. Hakansson, J. Axelsson, M. Palmblad, P. Hakansson. (2000) Mechanistic Studies of Multipole Storage Assisted Dissociation. *J Am Soc Mass Spectrometry*, **11**, 210-217.

[70] Sannes-Lowery K.A.; Hofstadler S.A. (2000) Characterization of Multipole Storage Assisted Dissociation: Implications for Electrospray Ionization Mass Spectrometry Characterization of Biomolecules. *JASMS*. **11**, 1-9.

[71] McFarland, M.A., Hendrickson, C.L., and Marshall, A.G. (2004) Ion "threshing": Collisionally activated dissociation in an external octopole ion trap by oscillation of an axial electric potential gradient. *Analytical Chemistry*, **76**(6), 1545-1549.

[72] Hofstadler, S.A., Sannes-Lowery, K.A., and Griffey, R.H. (1999) Infrared multiphoton dissociation in an external ion reservoir. *Anal Chem*, **71**, 2067-70.

[73] Hofstadler, S.A., Drader, J.J., Gaus, H., Hannis, J.C., and Sannes-Lowery, K.A.(2003) Alternative approaches to infrared multiphoton dissociation in an external ion reservoir. *JASMS*. **14**, 1413-23.

[74] Senko M.W.; Hendrickson C.L.; Emmett M.R.; Shi S.D.H.; Marshall A.G. (1997) External accumulation of Ions for Enhanced Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *JASMS*. **8**, 970-976.

[75] Sannes-Lowery, K. A.; Griffey, R. H.; Kruppa, G. H.; Speir, J. P.; Hofstadler, S. A. (1998) Multipole Storage Assisted Dissociation, a Novel In-Source Dissociation Technique for Electrospray Ionization Generated Ions. *Rapid Commun. Mass Spectrom*. **12**, 1957-1961.

[76] Belov, M.E., Gorshkov, M.V., Udseth, H.R., and Smith, R.D. (2001) Controlled ion fragmentation in a 2-D quadrupole ion trap for external ion accumulation in ESI FTICR mass spectrometry. *JASMS*. **12**,1312-1319.

[77] Pan, C., Hettich, R.L. (2005) Multipole-Storage Assisted Dissociation (MSAD) for the Characterization of Large ProteinsMixtures by ESI-FTICR-MS. *Anal. Chem*. **77**, 3072-3082.

[78] Keller, K.M., Brodbelt, J.S., Hettich, R.L., and Van Berkel, G.J.(2004) Comparison of sustained off-resonance irradiation collisionally activated dissociation and multipole storage-assisted dissociation for top-down protein analysis. J *Mass Spectrom*. **39**, 402-11.

[79] Sannes-Lowery K.A.; Hofstadler S.A. (2000) Characterization of Multipole Storage Assisted Dissociation: Implications for Electrospray Ionization Mass Spectrometry Characterization of Biomolecules. *JASMS*. **11**, 1-9.

[80] Palmblad, M., Hakansson, K., Hakansson, P., Feng, X.D., Cooper, H.J., Giannakopulos, A.E., Green, P.S., and Derrick, P.J.(2000) A 9.4 T Fourier transform ion cyclotron resonance mass spectrometer: description and performance. *Eur. J. Mass Spectrom*. **6**, 267-275.

[81] Haselmann K.F.; Bundnik B.A.; Kjeldsen F.; Nielsen M.L.; Olsen J.V.; Zubarev R.A. (2002) Electronic excitation gives informative fragmentation of polypeptide cations and anions. *Eur. J. Mass Spectrom*. **8**, 117-121.

[82] Kubinyi, H. (1991) Calculation of isotope distributions in mass spectrometry-A trivial solution for a nontrivial problem. *Anal. Chim. Acta*. **247**, 107–119.

[83] McDonald, W. H.; Tabb, D. L.; Sadygov, R. G.; MacCoss, M. J.; Venable, J.; Graumann, J.; Johnson, J. R.; Cociorva, D.; Yates, J. R. (2004) MS1, MS2, and SQT—Three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. *Rapid Commun. Mass Spectrom*. **8**, 2162–2168.

[84] PROWL: http://prowl.rockefeller.edu/

[85] Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. (2000) Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *JASMS*. **11**, 320–332.

[86] McDonnell, L.A., Giannakopulos, A.E., Derrick, P.J., Tsybin, Y.O., and Hakansson, P. (2002) A theoretical investigation of the kinetic energy of ions trapped in a radio-frequency hexapole ion trap. *Eur. J. Mass Spectrom*. **8**, 181-189.

[87] Uchiki, T., Hettich, R., Gupta, V., and Dealwis, C. (2002) Characterization of monomeric and dimeric forms of recombinant Sm1lp-histag protein by electrospray mass spectrometry. *Anal. Biochem*. **301**, 35-48.

[88] Link, A.J.; Eng, J.; Schieltz, D.M.; Carmack, E.; Mize, G.J; Morris, D.R.; Garvik, B.M.; Yates, J.R. 3rd (1999) Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.*, **17**, 676-682.

[89] Wittman ,H. G.(1982) Components of bacterial ribosomes. *Annu. Rev. Biochem*. **51**, 155-183.

[90] Arnold, R.J; Reilly, J.P. (1999) Observation of *Escherichia coli* ribosomal proteins and their posttranslational modifications by mass spectrometry. *Anal. Biochem.*, **269**, 105-112.

[91] R. M. Kamp, R. M.; Srinivasa, B.R.; Von Knoblauch, K.; Subramanian, A. R. (1987) Occurrence of a methylated protein in chloroplast ribosomes. *Biochemistry* **26**, 5866-5870.

[92] Yamaguchi, K.; Subramanian, A. R. (2000) Identification of all the proteins in the 50S subunit of an organelle ribosome (chloroplast) *J. Biol. Chem*. **275**, 28466-28482.

[93] Kowalak, J.A.; Walsh K.A. (1996) Beta-methylthio-aspartic acid: identification of a novel posttranslational modification in ribosomal protein S12 from Escherichia coli. *Protein Science*, **5**, 1625-1632.

[94] Arnold, R. J.; Polevoda, B.; Reilly, J. P.; F. Sherman, F.(1999) the action of N-terminal acetyltransferases on yeast ribosomal proteins. *J. Biol. Chem*. **274**, 37035-37040.

[95] Kalholz, B. P.; Myasnikov, A. G.; van Heel, M. (2004) Visualization of release of 3 on the ribosome during termination pf protein synthesis. *Nature*. **427**, 862-865.

[96] Javelle, E. Severis, J. Thornton, J. Merrick. (2004)Ammonium Sensing in *Escherichia coli*. *J. Bio. Chem*. **279,** 8530-8538.

[97] van Heeswijk, W. C.; Wen,D.; Clancy, P.; Jaggi, R.; Ollis, D. L.; Westerhoff, H. V.; Vasudevan, S. G. (2000) The *Escherichia coli* signal transducers PII (GlnB) and GlnK form heterotrimers *in vivo*: Fine tuning the nitrogen signal cascade. *PNAS*, **97**, 3942-3947.

[98] Arcondeguy, T.; Jack, R.; Merrick, M.(2001) P(II) Signal Transduction Proteins, Pivotal Players in Microbia Nitrogen Control, *Microbial Mol. Biol. Rev*. **65**, 80-105.

[99] Ninfa, A. J.; Atkinson, M. R. (2000) P(II) Signal Transduction Proteins, *Trends Microbiol*. **8**, 172-190.

[100] Thomas, G.; Coutts, G.; Merrick, M. (2001) The glnK amtB operon: a conserved gene pair in prokaryotes. *Trends in Genetics* **16**, 11-14.

[101] Atkinson, M. R.; Blauwkamp, T. A.; Ninfa, A. J. (2002) Context-Dependant Functions of the PII and GlnK Signal Transduction Proteins in *Escherichia coli*. *J. Bacteriol.* **184**, 5364-5375.

[102] Javelle and Merrick (2005) Complex formation between AmtB and GlnK: an ancestral role in prokaryotic nitrogen control. *Biochem. Soc. Trans*. **33**,170-172.

[103] Zheng, L.; Kostrewa, D.; Berneche, S.; Winkler, F. K.; Li, X. (2004) The mechanism of ammonia transport based on the crystal structure of AmtB and *Escherichia coli*. *PNAS* **101**, 17090-17095.

[104] Forchhammer, K.; Hedler, A.; Strobel, H.; Weiss, V. (1999) Heterotrimerization of PII-like signaling proteins: implications for PII-mediated signal transduction systems. *Molecular Micro.* **33**, 338-349.

[105] Atkinson, M. R.; Ninfa, A. J. (1999) Characterization of the GlnK protein of *Escherichia coli*, Mol. Microbiol. **32**, 301-313.

[106] Atkinson, M. R.; Ninfa, A. J. (1998) Role of the GlnK Signal Transduction Protein in the Regulation of Nitrogen Assimilation in *Escherichia coli*. *Mol. Microbiol*. **29**, 431-447.

[107] Blauwkamp, T.; Ninfa, A. J. (2002) Physiological Role of the GlnK Signal Transduction Protein of *Escherichia coli*: Survival of Nitrogen Starvation. *Mol. Microbiol.* **46**, 203-214.

[108] Coutts, G.; Thomas, G.; Blakey, D.; Merrick, M. (2002) Membrane Sequestration of the Signal Transduction Protein by the Ammonium Transporter AmtB. *EMBO J*. **21**, 536-545.

[109] Javelle, A.; Serveri, E.; Thornton, J.; Merrick, M. (2004) Ammonium Sensing in Escherichia coli. Role of the Ammonium Transproter AmtB and AmtB-GlnK complex Formation. *J Biol. Chem*. **279**, 8530-8538.

[110] Drepper, T.; Grob, S.; Yakunin, A. F.; Hallenbeck, P. C.; Masepohl, B.; Klipp, W. (2003) Role of GlnB and GlnK in ammonium control of both nitrogenase systems in the phototrophic bacterium *Rhodobacter capsulatus*. *Microbiology*. **149**, 2203-2212.

[111] Zhang, Y.; Pohlmann, E. L.; Ludden, P. W.; Roberts, G. P. (2001) Functional Characterization of Three Homologs in the Photosynthetic Bacterium *Rhodospirillum rubrum*: Roles in Sensing Ammonium and Energy Status. *J. Bac*. **183**, 6159-6168.

[112] Maheswaran, M.; Frochhammer, K. (2003) Carbon-Source-Dependent Nitrogen Regulation in Escherichia coli is mediated Through Glutamine-Dependent GlnB Signalling. *Microbiology*. **149,** 2163-2172.

[113] Perlova, O.; Ureta, A.; Nordlund, S.; Meletzus, D. (2003) Identificaiton of Three Genes Encoding PII-Like Proteins in *Gluconacetobacter diazotrophicus*: Studies of Their Role(s) in the Control of Nitrogen Fixation, *J. Bac*. **185**, 5854-5861.

[114] Mann, M.; Hendrickson, R. C.; Pandey, A. (2001) Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem*. **70**, 437-473.

[115] LeDuc, R. D.; Taylor, G.K; Kim, Y. B.; Januszyk, T. E.; Bynum, L. H.; Sola, J. V.; Garavelli J. S.; Kelleher, N. L (2004) ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry. *Nucleic Acids Research*. **32**, W340-W345.

[116] Holmes, M. R.; Giddings, M. C. (2004) Prediction of Posttranslational Modifications Using Intact-Protein Mass Spectrometric Data. *Anal. Chem*. **76**, 276-282

[117] Narasimhan, C.; Tabb, D. L.; VerBerkmoes, N. C.; Thompson, M. R.; Hettich, R. L.; Uberbacher, E. C. (2005) MASPIC: Intensity-based tandem mass spectrometry scoring scheme that improves peptide identification at high confidence. *Anal. Chem*. **77**, 7581–7593. http://compbio.ornl.gov/MASPIC/]

[118] Tabb, D.L.; Narasimhan, C.; Strader, M.B.; Hettich, R.L. (2005) DBDigger: Reorganized proteomic database identification that improves flexibility and speed. *Anal. Chem.* **77**, 2464-74.

[120] Hardy, S.J.S.; Kurland, C.G.; Voynow, P.; Mora, G. (1969) Ribosomal proteins of *Escherichia coli* .I. Purification of 30S ribosomal proteins. *Biochem.* ***8*****,** 2897.

[121] Verberkmoes, N.C., et al. (2006) Determination and comparison of the baseline proteomes of the versatile microbe *Rhodopseudomonas palustris* under its major metabolic states. *J. Proteome Research*. **5,** 287-298.

[122] Dixon, R.; Kahn D. (2004) Genetic regulation of biological nitrogen fixation. *Nat. Rev. Micro*. **2**, 621-631.

[123] Connelly, H.M.; Pelletier, D.A.; Tse-Yuan, Lankford, P.K.; and Hettich, R.L. (2006) Characterization of GlnK and GlnB Uridylylation in Response to Nitrogen Availability for *Rhodopseudomonas palustris*. *Analytical Biochemistry, Accepted, In Press.*

[124] Bendtsen J.D.; Nielsen H.; von Heijne G.; Brunak S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*. **340**, 783-95.

[125] Hiller K.; Grote A.; Scheer M.; Munch R.; Jahn D. (2004) PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Res*. **32**, W375-9.

[126] Gardy JL, Laird MR, Chen F, Rey S, Walsh CJ, Ester M, Brinkman FS. (2004) PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics*. **21**, 617-23.

Vita

Heather Marie Connelly was born in Atlanta, Georgia on Decermber 5th 1977. She grew up in the small town of Jasper, Georgia with her parents, sister, and brother, where she graduated from Pickens County High School, as an honor student, in 1996. She received her Associate of Natural Science from Reinhardt College in 1998 and further received a Bacelor of Science degree in Chemistry and Biology from Shorter College in 2000. Heather pursued and recived a Master of Science degree in Biology from the State University of West Georgia under the direction of Dr. Leos Kral in 2002. Her thesis work was entitled the "Genetic Population Structure of the Tallapoosa Shiner".

She enrolled in the University of Tennessee-Oak Ridge National Laboratory Graduate School of Genome Science and Technology in 2002 to pursue her doctorate in Life Sciences. She graduated with a Ph.D. in 2006. She moved to Atlanta in 2006 and is pursuing research options.